ARDIŞIK TEKRARLI DNA DİZİLERİNİN OPTİMUM DÜZEYDE BULUNMASINA YÖNELİK PROGRAMLAMA ÇALIŞMASI

Pamukkale Üniversitesi
Fen Bilimleri Enstitüsü
Yüksek Lisans Tezi
Elektrik-Elektronik Mühendisliği Ana Bilim Dalı

Onur İNAN

Danışmanlar: Prof. Dr. Mustafa TEMİZ, Yrd. Doç. Dr. A. Kadir YALDIR

Ağustos 2006
DENİZLİ

YÜKSEK LİSANS TEZİ ONAY FORMU

Onur İNAN tarafından Prof. Dr. Mustafa TEMİZ ve Yrd. Doç. Dr. A. Kadir YALDIR yönetiminde hazırlanan "**Ardışık Tekrarlı DNA Dizilerinin Optimum Düzeyde Bulunmasına Yönelik Programlama Çalışması**" başlıklı tez tarafımızdan okunmuş, kapsamı ve niteliği açısından bir Yüksek Lisans Tezi olarak kabul edilmiştir.

Jüri Başkanı (1. Danışman)
Prof. Dr. Mustafa TEMİZ

Jüri Üyesi (2. Danışman)
Yrd. Doç. Dr. A. Kadir YALDIR

Jüri Üyesi
Yrd. Doç. Dr. Ahmet ÖZEK

Jüri Üyesi
Yrd. Doç. Dr. Murat AYDOS

Jüri Üyesi
Yrd. Doç. Dr. Sezai TOKAT

Pamukkale Üniversitesi Fen Bilimleri Enstitüsü Yönetim Kurulunun
..../..../..... tarih ve ..... sayılı kararı ile onaylanmıştır.

Prof. Dr. Mehmet Ali SARIGÖL
Müdür

Bu tezin tasarımı, hazırlanması, yürütülmesi, araştırılmalarının yapılması ve bulgularının analizlerinde bilimsel etiğe ve akademik kurallara özenle riayet edildiğini; bu çalışmanın doğrudan birincil ürünü olmayan bulguların, verilerin ve materyallerin bilimsel etiğe uygun olarak kaynak gösterildiğini ve alıntı yapılan çalışmalara atfedildiğini beyan ederim.

İmza: 

Öğrenci Adı Soyadı: Onur İNAN

TEŞEKKÜR

Tez çalışmam boyunca sabırla bana destek olup, yardımlarını benden esirgemeyen, tez danışmanlarım ve değerli hocalarım Prof. Dr. Mustafa TEMİZ'e ve Yrd. Doç. Dr. A. Kadir YALDIR'a, önerilerinden dolayı araştırma görevlisi arkadaşlarıma, bilgi ve tecrübelerinden faydalandığım değerli meslektaşım Önder Çivril'e, üzerimde emeği geçen tüm Mühendislik Fakültesi hocalarıma, Denizli'de manevi desteklerini ve dostluklarını esirgemeyen değerli arkadaşlarıma; Denizli'ye gitmem konusunda esneklik gösteren ve desteğini esirgemeyen Bucak Emin Gülmez MYO Müdür Yardımcı A. Alper Yarış'a, manevi desteklerinden ötürü Bucak Emin Gülmez MYO akademik ve idari personeline, çalışmalarımda destekten ziyade engel olan ve her türlü vaktimi almalarına rağmen bana bu mesleği sevdiren, moral motivasyon kazandıran öğrencilerime, İngilizce literatür çalışmalarında hatalı ve eksik çevirilerimi düzeltmede yardımcı olan değerli öğretmen arkadaşım Celil Özkılınç'a gönülden teşekkürü borç biliyorum.

Tez konumu belirlemede yardımlarından ötürü Doç. Dr. Mehmet Karaca'ya; manevi destekleri, yardımları ve sonsuz sevgilerinden ötürü beni bugünlere getiren üzerimde haklarını ödeyemeyeceğim en büyük emekleri bulunan değerli aileme ve adını şu an yazamadığım bana emeği geçmiş herkese teşekkürlerimi sunuyorum.

ÖZET

ARDIŞIK TEKRARLI DNA DİZİLERİNİN OPTİMUM DÜZEYDE BULUNMASINA YÖNELİK PROGRAMLAMA ÇALIŞMASI

İnan, Onur

Yüksek Lisans Tezi, Elektrik-Elektronik Mühendisliği ABD

Tez Yöneticileri: Prof. Dr. Mustafa Temiz, Yrd. Doç. Dr. A. Kadir YALDIR

Haziran 2006, 62 Sayfa

Deoksiriboz nükleik asit (DNA) de bulunan ardışık tekrarlar iki veya daha fazla nükleotid motifinin ardışık, birbirine benzer kopyalarıdır. Ardışık tekrarların hastalıklara neden olduğu, düzenleyici ve evrimsel roller oynayabildiği ve önemli bir laboratuvar ve analitik araç olduğu gözlemlenmiştir. Mini uydular veya basit ardışık tekrarlarında (Simple Sequence Repeat – SSR) görüldüğü gibi ardışık tekrarların DNA üzerinde yerini gösteren işaretleri (markır) olarak kullanılabilmeleri pek çok araştırıcının ilgisini çekmiştir. DNA markırları genetik analizlerin hızını artırarak genetik biliminde devrime yol açmıştır.

Basit ardışık tekrarlar (SSR) 1 ile 5 baz uzunluğunda nükleotid motiflerinin tekrar etmesidir ve genomda bol miktarda bulunuşları, aşırı değişken yapıları ve yüksek çıktılı analizlere uygunluğu bakımından günümüzde pek çok bitki ve hayvan genomlarında tercih edilen markırlardır. SSR ler bir kez geliştirildikten sonra son derece değerlidirler. Fakat elde edilmeleri zaman alıcı, pahalı ve aşırı işgücü gerektirir. Pek çok genoma ait diziler kamuya açık veri bankalarından ücretsiz elde edilebilirler ve hesaplama yöntemlerinin kullanılmasıyla bu kaynakların taranması sonucu markır geliştirilmesi hızlı ve ekonomik olur. İfade edilmiş ardışık etiketler (Expressed Sequence Tags – EST) sadece bol miktarda elde edilebilmeleri yüzünden değil; aynı zamanda ifade edilmiş genleri de temsil ettikleri için SSR leri bulmada ideal adaylardır. Ardışık tekrarların motif boyutları, kopya sayıları, mutasyon geçmişleri vs hakkında ayrıntılı bilgiler edinmek mevcut algoritmaların bazı yetersizlikleri nedeni ile sınırlıdır.

Bu çalışmada, Tandem Repeats Miner adı verilen motif ve motif boyutları verilmesine gerek duyulmadan çalışan yeni bir yazılım sunulmuştur. Gen bankasından değişik özelliklere sahip bazı diziler seçilerek dizi koleksiyonu oluşturuldu. Bu koleksiyonu yapmanın ana amacı, geliştirilen algoritmanın geçerliliğini doğrulamak için referans noktaları oluşturmaktır. Bu diziler, DNA dizilerinde karşılaşılan ardışık tekrar bölgelerindeki ortak problemler için bir altyapı sağlamaktadır. Bu koleksiyondaki diziler kullanılarak Tandem Repeats Miner algoritmasının sonuçları, Tandem Repeat

Finder ve Hauth algoritması gibi popüler algoritmalarla karşılaştırılmaktadır. Tandem Repeats Miner DNA dizilerindeki VNTR (Variable Number Tandem Repeats – Değişken Sayıda Ardışık Tekrarlar) ve SSR bölgelerini başarıyla belirlemektedir. Analiz sonucu ardışık tekrar bölgelerinin motif boyutunun, kopya sayısının geniş bir aralığı kapsadığı ve karmaşık motif yapıları gösterdiği belirlenmiştir.

Anahtar Kelimeler: Ardışık Tekrar, SSR, EST, Gen Bankası

Prof. Dr. Mustafa TEMİZ
Yrd. Doç. Dr. A. Kadir YALDIR
Yrd. Doç. Dr. Ahmet ÖZEK
Yrd. Doç. Dr. Murat AYDOS
Yrd. Doç. Dr. Sezai TOKAT

ABSTRACT

PROGRAMMING ON FINDING TANDEM REPEAT DNA SEQUENCES AT OPTIMUM LEVEL

İnan, Onur

M. Sc. Thesis in Electrical&Electronics Engineering

Supervisors: Prof. Dr. Mustafa Temiz, Asst. Prof. Dr. A. Kadir YALDIR

June 2006, 62 Pages

A tandem repeat in DNA is two or more contiguous, approximate copies of a motif of nucleotides. Tandem repeats which have been shown to cause human disease, may play a variety of regulatory and evolutionary roles, and are important laboratory and analytic tools. Repeats containing DNA sequences have attracted many researches since their use in DNA marker technologies, such as microsatellities or simple sequence repeats (SSRs). DNA markers have revolutionized the field of genetics by increasing the pace of genetic analysis.

Simple sequence repeats (SSRs) are repetitions of nucleotide motifs of 1 to 5 bases and are currently the markers of choice in many plant and animal genomes due to their abundant distribution in the genomes, hyper variable nature and suitability for high-throughput analysis. While SSRs, once developed, are extremely valuable, their development is time consuming, laborious and expensive. Sequences from many genomes are continuously made freely available in the public databases and mining of these sources using computational approaches permits rapid and economical marker development. Expressed Sequence Tags (ESTs) are ideal candidates for mining SSRs not only because of their availability in large numbers but also due to the fact that they represent expressed genes. Extensive knowledge about motif size, copy number, mutational history, etc, for tandem repeats has been limited by the inability to easily detect them in genomic sequence data.

In this study, a new software is called Tandem Repeats Miner presented, for finding tandem repeats which works without the need to specify either the motif or motif size. A collection of GenBank sequences is constituted representing tandem repeat regions having simple and complex motif structures. The purpose of the sequence collection is to provide a benchmark for validating the identification algorithm. These sequences provide the framework for common problems encountered in tandem repeat regions in DNA sequences. Using these GenBank sequences, the results of Tandem Repeats Miner

is compared with popular algorithms such as Tandem Repeat Finder and Hauth's algorithm. Tandem Repeats Miner successfully identifies the SSR regions and VNTR (Variable Number Tandem Repeats) regions in DNA sequences. The analysis determined that tandem repeat regions cover a wide range of motif sizes, copy numbers and exhibit complex motif structures.

Keywords: Tandem Repeat, SSR, EST, GenBank

Prof. Dr. Mustafa TEMİZ
Asst. Prof. Dr. A. Kadir YALDIR
Asst. Prof. Dr. Ahmet ÖZEK
Asst. Prof. Dr. Murat AYDOS
Asst. Prof. Dr. Sezai TOKAT

İÇİNDEKİLER

Sayfa

ŞEKİLLER DİZİNİ

Sayfa

TABLOLAR DİZİNİ

Sayfa

SİMGE VE KISALTMALAR DİZİNİ

bp	baz çifti
DNA	Deoksiriboz nükleik asit
DP	Dinamik programlama
EST	İfade edilmiş ardışık etiketler
mRNA	Haberci RNA
RNA	Riboz nükleik asit
SNP	Tek nükleotidli polimorfizim
SSR	Basit ardışık tekrarlar
TR	Ardışık Tekrarlar
VLTR	Değişken uzunlukta ardışık tekrarlar
WDP	Sarmal dinamik programlama

1. GİRİŞ

Yirminci yüzyılın son çeyreğinden itibaren DNA teknolojilerindeki yeni gelişmeler (DNA chips, protein chips, PCR chips ve otomasyonlar) Genetik Mühendisliği, Biyomühendislik, Farmakogenetik, Biyoinformatik ve Proteomik kavramlarının oluşumunu sağlamıştır. DNA, RNA, protein yapı ve fonksiyonlarının incelenmesinde, yeni genlerin bulunmasında veya kopyalama değişikliklerinin (transkripsiyon varyantları) ve çok biçimliliklerinin (polimorfizmler) belirlenmesinde etkin bir şekilde kullanılan biyoinformatiksel yaklaşımlar günümüzde Biyoinformatik bilim dalını oldukça önemli bir konuma getirmiştir.

Yukarıda belirttiğimiz yeni gelişmeler, araştırıcıların oldukça fazla oranda DNA dizi (sekans) verileri elde edebilmelerine olanak sağlamıştır. DNA dizilerinin belirlenmesinde (sekanslama) robotik otomasyonun kullanılmasıyla bitki, hayvan ve diğer organizmalara ait milyonlarca DNA Sekansı Gen Bankalarında toplanmıştır.

Bu gelişmelerden hareketle ulaşılmak istenen hedef; Gen Bankası verilerini mikro uydu ve mini uydu içerikleri yönünden Ardışık Tekrarlı DNA dizilerinin dağılımları ve fonksiyonları hakkında yeterli bilgiyi verebilen bir yazılımın geliştirilmesi olmuştur.

Ardışık Tekrarlı DNA dizilerinin bu derece önem arz etmesinin nedeni; son yıllarda yapılan çalışmalarda bu dizilerdeki değişikliklerin özellikle insanda görülen sinir sistemi ile ilgili hastalıklarda etkin olduğunun, diğer bazı organizmalarda da gen ifadesinde yer aldığının ve bazı durumlarda ise kodladığı protein üzerinde önemli etkileri olduğunun gözlenmiş olmasıdır.

Genomların (kromozom topluluğu) hem kodlanan hem de kodlanamayan bölgelerinin en ilginç özelliği kısa ardışık tekrarlı DNA dizilerini içermesidir (Dizi = Sequence).

Bunlar Tandem Repeats (TR) diye adlandırılır. Geliştirilen analiz yazılımında dizi motiflerinin (motif – kalıp ACTGGGA gibi) ayrıntılı taranması sonucu Tam, Tam olmayan ve Birleşik TR'ler ayrı ayrı anahtar kelimeler ile tesbit edilmeye çalışılmaktadır. Analiz yazılımında, üzerinde araştırma yapılacak gen verisi dosyaları FASTA (dizileri karşılaştırmak için geliştirilmiş dizilim yazılımı) formatında ele alınmaktadır. Kullanıcı tanımlamalarına ya da seçeneklerine bağlı olarak değişken motif uzunluklarında eş zamanlı olarak değişen motif uzunluklarının taranması işlemi, analiz yazılımının arama modülünde yer almaktadır.

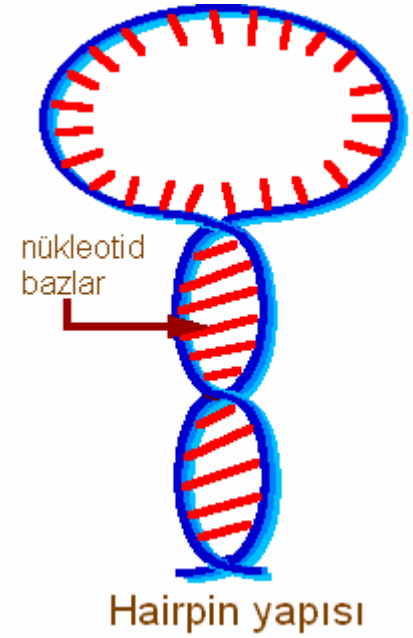
"Tandem Repeats Miner" adı verilen analiz yazılımında tarama sonuçlarının, kullanıcı tarafından geliştirilebilecek sonuçlar içermesine dikkat edilmiştir. Özellikle TR'lerin EST'lerde (EST- Expressed Sequence Tag) tespiti, değişik çevre koşullarında, baskı durumundaki, gelişme dönemindeki organlarda ve dokularda yeralan önemli genlerin kodlama bölgelerindeki TR'lerin keşfedilmesinde ve gen haritalarının çıkarılmasında da önemli katkılar sağlayacaktır.

1.1 Ardışık Tekrarlı DNA Dizilerinin Önemi

Daha önce yapılan analiz yazılımları, şu sonuçları ortaya çıkarmıştır: Organlarda, dokularda ve gelişme dönemlerinde gözlenen ardışık tekrar ve ardışık tekrar birleşim sayıları farklıdır. Bu da göstermektedir ki; ardışık tekrarların dokularda ve organlarda dağılımı rastgele değildir. Transkripsiyonu alınmamış diğer tekrar dizilerinden farklılık göstermektedir.

Ardışık Tekrarlar (TR) olarak adlandırılan DNA dizilerinin hem çekirdeği olmayan canlılarda (prokaryotlar) hem de çekirdeği olan canlılar da (ökaryotlar) gözlemlendiği, genomlarda rastgele dağıldığı ifade edilmiştir (Jeffreys vd 1985, Heslop – Harrison 2003). TR'lerin bazıları genlerin düzenlenmesinde önemli rol oynamaktadırlar; bazılarının ise herhangi bir fonksiyonu yoktur. Buna rağmen her birinin DNA gösterimi ve genetik bağlantı analizi açısından ne derece önemli olduğu açıktır (Scott vd 2000, Toth vd 2000). DNA dizilerindeki tekrarlar şu nedenlerle araştırıcıları ilgilendirir:

1. Bazı yapısal veya kopyalama (replication) mekanizmalarında saç tokası (hairpin) yapılarının oluşumunda önemli role sahiptir (MC Murray vd 1999, Keniry vd 2000, Shafer ve Smirnov 2000). Hairpin yapısı, Şekil 1.1 de verilmektedir.



Şekil 1.1 Kopyalama aşamasındaki Hairpin yapısı

2. Artan sayıda sinirsel düzensizlikler, ardışık tekrarlı DNA dizileri ile ilişkilendirilmiştir (Reddy ve Housman 1997, Timchenko ve Caskey 1999).
3. DNA işaretleyici (markır) teknolojilerinde kullanımları:
   - Mikro uydu ve basit ardışık tekrarlar (SSRs)
   - Ara basit ardışık tekrarları (ISSRs) ve minisatellitlerin markır yardımlı seçimde (MAS) doğrudan kuvvetlendirilmesi (DAMD-PCR)
   - Konuma bağlı klonlama (positional cloning)
   - Miktara ve niteliğe bağlı konumların tanımlanması, soya ve gelişime ait gen haritalarının çıkarılması (Scott vd 2000, Karaca vd 2002).

   Bu üç nedenle birlikte son gelişmeler göstermektedir ki, bazı değişken sayıdaki TR'ler (VLTR) ve SSR dizileri;
   - Transkripsiyonun (kopyalama) düzenlenmesinde
   - mRNA'ların etkinliğinin veya kararlığının tespit edilmesi ya da proteinlerin yapısının değiştirilmesi suretiyle aktivitelerinin modifiye edilmesi (değiştirilmesi) hususlarında da önemli bir rol üstlenmektedir.

1.2 DNA Dizileriyle İlgili Genel Kavramlar

EST'ler tek geçişli DNA dizileridir. 200-500 nükleotit uzunluğunda olup herhangi bir dokuda veya gelişme döneminde ifade edilen genleri temsil eden haberci RNA (mRNA) ya da tamamlayıcı DNA (cDNA – complementary DNA)'dan elde edilirler. Tipik bir EST; gen transkripsiyonunun (belirtilmiş ya da belirtilmemiş klonlama) kodlama bölgesinin tek bir parçasını içerir. EST'nin faydalarından birisi de; organ, doku ve gelişme dönemine ait bir pattern'in (motifin) nükleotit yapısını ortaya çıkarmasıdır.

Dokulardaki özelleşmiş EST populasyonlarının birleşimi bu yüzden ifade edilmiş genlere ayrıntılı bir bakış açısı verir ve netice itibariyle fiziksel davranışların kökenindeki biyokimyasal yolların anlaşılmasında ve genlerin keşfedilmesinde yeni bir yöntem olarak kullanılmaktadır. EST'ler tek nükleotit çok çeşitliliğin (single nucleotide polymorphism - SNP) araştırılıp ortaya çıkarılması amacıyla kullanılırlar (Schmidt 2003). Ayrıca basit ardışık tekrarları (SSR) için de faydalanılmaktadır (Thiel vd 2003).

Basit ardışık tekrarları (SSR) 1-6 bp motif uzunluğundaki basit ardışık tekrarlardan oluşan DNA kısımlarını temsil etmektedir. SSR'ler ideal DNA işaretleyicileridir çünkü bireyler açısından oldukça çokbiçimlidir (polymorphic) ve genomlar arasında bolca dağılmıştır (Klintschar ve Wiegand 2003). SSR'ler ayrıca kalıtımsal olarak elde edilebilmektedir. Bu sayede TR'leri yandan kuşatan tekil primer çiftlerinin kullanılması sayesinde, laboratuarda genomları çoğaltmak amacıyla kullanılan PCR cihazı tarafından hızlı ve kolay bir şekilde tespit edilebilirler. Bunun da ötesinde, genetik ve fiziksel haritaların çıkarılmasında sekans (dizi) belirleyici bir rol üstlenmektedir (Karaca vd 2002).

SSR'leri geliştirmek için genel prosedür şu şekildedir: Küçük girişli genom kütüphanelerinin kurulumu; bunu takiben ardışık tekrarlı oligonükleotitler ile hibritleştirme ve tek bir hücreden elde edilen genetik olarak bağlantılı bir grup hücre veya organizmayı temsil eden klonların bir dizi içinde düzenlenmesi. Böylece işlem zamanı hem kısaltılacaktır; hem de çalışma yoğunluklu bir işlem haline getirilecektir. SSR'lerin gelişiminde alternatif strateji de; artan miktarda gen bilgisinin genomik DNA ve EST veri tabanlarından temin edilmesidir. Dizi bilgisindeki hızlı artışa bağlı olarak EST-SSR'lerin üretimi, mevcut genomik SSR'lere göre cazip bir alternatif haline

gelmektedir (Thiel vd 2003). SSR primer çiftlerinin gelişiminin önemli miktarda azalan maliyetlerde olması, EST-SSR'lerin halen büyümekte olan EST veritabanlarından serbestçe temin edilmesinden kaynaklanmaktadır. EST'ler genomların transkripsiyona uğramış kısmını temsil ettiklerinden, EST-SSR işaretleyicileri gen haritalarının doğrudan çıkarılmasına katkıda bulunurlar.

SSR'ler genlerin önemli kodlama bölgelerinde yer almaktadır. Bu kodlama bölgeleri çevrenin, baskı durumunun, organların, dokuların ve gelişme aşamalarının çeşitli durumlarını ifade eder ve organlardaki, dokulardaki ve gelişme aşamalarındaki özel SSR'lerin gelişimine katkıda bulunur. Böylece genlerdeki tekrar fonksiyonları, organizmaların soy haritalarının çıkarılması ve diğer ileri çalışmalar daha anlaşılır hale gelmiştir.

Ardışık tekrarlı DNA dizilerini belirlemede, Sputnik (Abajian 1994); Tandem Repeats Finder (TRF) (Benson 1999); REPuter (Kurtz vd 2001); Simple Sequence Repeat Identification Tool (SSRIT) (Kantety vd 2002); FindPatterns; Simple Sequence Repeat Finder (SSRF) (Sreenu vd 2003); Repeat Finder; STRING (Parisi vd 2003); Microsatellite Search (MISA) (Thiel vd 2003); Tandem Repeats Analyzer (TRA) (Bilgen vd 2004) gibi birkaç yazılım geliştirilmiştir. Ardışık tekrarlı DNA dizilerini belirlemede kullanılan bu yazılımlar çok faydalı olmalarına rağmen, uzunluğu sınırlı dizilerde çalışabilmeleri, tam olmayan ve/veya birleşik ardışık tekrarları bulamama gibi kullanımlarını sınırlayan birçok dezavantaja sahiptirler. Bu yazılımlardan bazıları ileride karşılaştırılacaktır.

2. DNA DİZİLERİNİN ELDE EDİLMESİ VE KULLANIM ALANLARI

Genetik biliminde ilk olarak fenotipik veya morfolojik markırlar (işaretleyiciler) ve sonraları isozyme (protein) markırları yüzyılı aşkın süredir yoğun olarak kullanılmalarına rağmen, ancak 20. YY ikinci yarısından itibaren DNA markırlarının etkin olması ile genetik analizlerin doğruluğu ve gelişim hızı artmıştır (Dodgson vd 1997). DNA markırlarının daha etkin bir konuma gelmesi, pek çok bitki ve hayvan genomlarında genom bağlantı (linkage) haritalarının çıkarılmasına, tarımsal ürünlerde gen klonlanmasına, genom analizi ve markırlara dayalı seleksiyon yöntemlerinin gelişmesine yol açmıştır (Cullis 2002, Dodgson vd 1997, Paterson 1996a). Böylece tarım konusunda çalışan araştırıcıların, tüketilebilir aşı üretimi, agronomik genlerin klonlanması, hastalık ve zararlılara karşı dirençli bitkilerin geliştirilmesi, hem verim hem de kalite yönünden üstün niteliklere sahip bitkilerin üretilmesi gibi geleneksel ıslah alanında gerçekleşmesi olanaksız görülen konularda başarılı çalışmalar yapmaları mümkün olmuştur. DNA markır teknolojileri, diğer taraftan, genetik teşhis, populasyon çalışmaları, karşılaştırmalı genomics, farmakogenomics, ilaç keşfi ve moleküler evrim, tıp ve adli vakaların açıklanmasında da giderek artan oranlar da kullanılmaya başlanmıştır (Bennetzen vd 1997, McCarthy ve Hilfiker 2000, Pfost vd 2000, Rafalski ve Tingey 1993, Terauchi ve Konuma 1994).

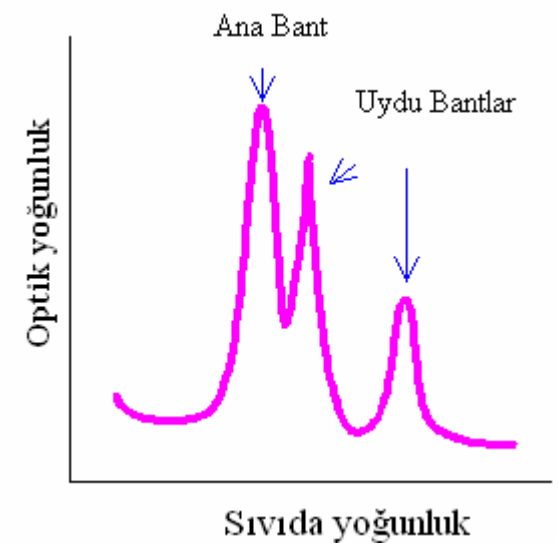
2.1 Moleküler Markır Teknolojisi ve Basit Dizi Tekrarları

İlk DNA markırları, sınırlandırılmış kısım uzunluk polimorfizmi (RFLP – Restriction Fragment Length Polymorphism) markırları, çok faydalı oldukları kanıtlanmasına rağmen bu markırların geliştirilmesi ve kullanımı işgücünü artırdığı, zaman alıcı olduğu, pahalı ve yüksek çıktılı otomasyona uygun olmadığı için hemen hemen kullanımdan kalkmıştır (Paterson 1996b, Rafalski ve Tingey 1993, Pfost vd 2000, Terauchi ve

Konuma 1994). Bu nedenlerle, rastlantısal çoğaltılmış DNA polimorfizmi (RAPD – Random Amplified Polymorphic DNA), çoğaltılmış kısım uzunluk polimorfizmi (AFLP – Amplified Fragment Lentgh Polymorphism), basit tekrar sekansları (SSRs = simple sequence repeats) gibi PCR'a dayalı markırlar, moleküler genetik çalışmaları için zaman içinde giderek daha çok popüler olmuştur (Paterson 1996b).

2.2 Ardışık Tekrarlar (TR) , İfade Edilmiş Ardışık Etiketler (EST) ve Basit Ardışık Tekrarları (SSR)

Ardışık Tekrarlar (TR) nükleotit dizilerinin ard arda sıralanmasıdır. Üç alt sınıfa ayrılırlar : uydular, mini uydular ve mikro uydular. Uydu adı ışık tayfından gelir. Şekil 2.1'de ardışık tekrarların sıvı ortamda, yoğunluklarına göre bant dağılımları gösterilmektedir.



Şekil 2.1 Uydu bantların resimlenmesi. Sıvı yoğunluk santrifüjü kullanıldığında, baz bileşimi önemli farklılık gösteren DNA parcaçıkları ayrılır ve sonra ultraviole ışık tayfına göre izlenir. Ana band DNA karışımını, uydu bantlar ise ardışık tekrarları gösterir.

2.2.1 Uydular

DNA uydularının boyutları 100 Kb (Kilobaz) ile 1 Mb (Megabaz) arasında değişir. İnsanların ve diğer organizmaların çoğu sentromerde bulunmaktadır.

2.2.2 Mini uydular

Mini uyduların boyutları 1 Kb ile 20 Kb arasında değişir. Mini uyduların en tanınmışı değişken sayılı ardışık tekrarlardır (VNTR). Bunların tekrarlayan birimleri 9 baz çifti (bp) ile 80 baz çifti arasında değişir. DNA'nın kodlama yapmayan bölgelerinde bulunmaktadır. Bir mini uydudaki tekrar sayısı bireyden bireye değişir. Bu özellik DNA parmak izinin temelini teşkil eder. Diğer bir tip mini uydu da telemor olarak adlandırılan kromozom uçlarında bulunur. İnsan germ hücrelerinde telemor boyutları yaklaşık 15 kb dır. Yaşlı somatik hücrelerde telemorun boyu kısalır. Telemor ardışık tekrarlanan **GGGTTA** dizisini içerir.

2.2.3 Mikro uydular

Mikro uydular aynı zamanda basit ardışık tekrarlar (SSR-short tandem repeats) olarak da bilinir. Çünkü tekrarlanan birim sadece 1 ilâ 6 bp arasında değişir ve tüm tekrarlanan bölümün uzunluğu 150 bp den daha azdır. Mini uydulara benzer şekilde belirli bir tekrarın sayısı bireyden bireye değişir. PCR'a dayalı markırlar içinde SSR markırları, analizleri için az miktarda genomik DNA örneği gerektirmesi ve yüksek çıktılı otomasyona uygunluğu açısından giderek daha çok popüler olmaktadır (Hearne vd 1992). SSR'leri çok popüler yapan iki özelliği ise bugüne kadar yapılan çalışmalarda bütün genom boyunca dağılmış olarak bulunmaları ve aşırı değişken (hypervariable) doğalarıdır (Powell vd 1996, Tautz ve Renz, 1984, Toth vd 2000). Örnek olarak, (Cardle vd 2000) bitki genomlarında yaklaşık olarak her 6 Kb (Kilobaz) da bir oranında, bol miktarda SSR bulunduğunu rapor etmektedirler. Aşırı değişkenlik terimi ile SSR dizisindeki tekrar sayısının bireyden bireye veya genotipten genotipe değiştiği anlatılmak istenmiştir. Bu özellik SSR'leri genetik markır olarak olağanüstü değerli kılmaktadır. İki, üç ve dört nükleotitli basit ardışık tekrarları (SSR'ler) çoğunlukla çekirdek genomlarının bağlantı haritalarının oluşturulmalarında kullanılmalarına rağmen tek nükleotitli basit tekrar sekansları kloroplast genomlarının populasyon genetik analizlerinde kullanılmaktadır (Powell vd 1995). SSR'ler PCR teknolojisi kullanılarak belirlenebildiğinden, moleküler genetik bağlantı (Morgante ve Olivieri 1993) ve populasyon (Powell vd 1995) çalışmaları için de yüksek çıktılı platformlar kullanılarak taranabilirler. İnsanlarda, üç nükleotitli SSR'lerin ondörtten daha çok kalıtsal nörodejeneratif hastalıkla ilişkisi olduğu bulunmuştur ve bu vakalarda SSR'lerin

belirlenmesi hastalığın teşhisi amacı ile kullanılmaktadır (Bryant-Greenwood 2002, Sinden vd 2002). Tablo 2.1'de bu kalıtsal hastalıklar gösterilmektedir.

SSR'lerin çok değerli olduğu ve giderek popüleritesinin arttığı bir diğer alan "genomics" dir ve bu alanda bir türden geliştirilen SSR'ler bu türe yakın veya uzak türlerin genetik haritalaması, karakterizasyonu, gen klonlanması, farklılık ve evrim çalışmalarında kullanılmaktadır (Cordeiro vd 2001, Eujayl vd 2001, Killian vd 1997, Moore vd 1991, Peakall vd 1998, Rallo vd 2003, Westman ve Kresovich 1998). Bu yaklaşım bitki genomik çalışmalarında son birkaç yıldır ivme kazanmıştır. Yapılan gözlemlerde bitkilerin genom boyutları büyük farklılıklar göstermesine rağmen, bitkilerin hem gen içeriğini hem de gen sıralamasını büyük oranda koruduğu bulunmuştur (Bennetzen ve Freeling 1993). Karşılaştırmalı genetik analizlerde de farklı bitki türlerinin çok benzer fonksiyonlar için genellikle aynı homolog genleri kullandığı bulunmuştur (Ahn vd 1993, Bennetzen ve Freeling 1993).

2.2.4 İfade edilmiş ardışık etiketler (EST- Expressed Sequence Tags)

EST'ler DNA dizilerinin küçük parçacıklarıdır ve 200 ile 500 nükleotit uzunluğundadır. Tamamlayıcı DNA'dan (cDNA), ifade edilmiş genin bir veya her iki ucundan dizilerin tekrar çoğaltılması ile elde edilir. EST'ler gen yerini belirlemede gerekli zamanı oldukça düşürdüğü için bilinen genleri avlamada çok güçlü araçlardır. Bu da içerdikleri SSR aracılığı ile olmaktadır. EST'leri kullanarak, bilim adamları Alzhemier hastalığı ve bazı kanser türlerinden bazı genleri hızla izole etmişlerdir (WEB_1 2006).

2.3 Bioinformatik Yaklaşımlar

Pek çok genoma ait nükleotit dizilimleri kamuya açık veri bankalarından kolaylıkla elde edilebildiğinden biyoinformatik yaklaşımlar hızla moleküler markır geliştirilmesine yönelmiştir. Veri bankası kaynakları uygun hesaplama algoritmaları yardımıyla SSR'leri bulmak için kullanılmaktadır. Nükleotit dizilimleri hakkında bilgi edinmek için gerekli olan zengin kütüphane oluşturulması gibi çok pahalı yatırımlara ihtiyaç duyulmamaktadır. Böylece sadece markır geliştirme maliyeti azaltılmış olmaz, aynı

Tablo 2.1 Üç nükleotitli ardışık tekrarlarla ilişkili degeneratif insan hastalıkları (Pearson ve Sinden 1998, Baldi vd 1999)

Hastalık	Patern	Kopya Sayısı (Sinden 1999)			Kopya sayısı (Baldi 1999)		
		Normal	Stabil olmayan	Etkilenmiş	Normal	Stabil olmayan	Etkilenmiş
Spinobulbar muscular atrophy (Kenedi hastalığı)	CAG	14-32		40-55	9-36	> 47	28-62
Huntington hastalığı	CAG	10-34	36-39	40-121	6-35	> 35	36-121
Spinocerebellar ataxia 1	6-39		40-81	6-35			40-81
Spinocerebellar ataxia 2	CAG	14-31		34-59	14-32		33-77
Spinocerebellar ataxia 3 (Machado Joseph hastalığı)	CAG	13-44		60-84	12-40		67-82
Spinocerebellar ataxia 6	CAG	4-18		21-28	4-17		20-30
Spinocerebellar ataxia 7	CAG	7-17		38-130	7-17		38-130
Dentadorubropallidoluysian atropy Haw River sendromu	CAG				3-36		49-88
Spastik paraplegia	CAG						
FRA16A	CCG	16-49		1000-1900			
Jacobsen sendromu	CCG	11	80	100-1000			
Myotonic dystrophy	CTG	5-37	50-80	80-3000	5-30	36-50	50 ilâ > 700
X-A sendromu	CGG	6-52	59-230	230-2000	5-52		200 ilâ > 1000
X-E sendromu	CCG	4-39	31-61	200-900	5-30	36-50	200 ilâ > 1000
X-F sendromu	CGG	7-40		36-1008			50 ilâ > 700
Friedreich'in ataksiası	GAA	6-29	>34-40	200-900	7-22	34-65	200 ilâ > 1000

zamanda çok kısa bir zaman diliminde çok sayıda markır geliştirilmesi mümkün olur. Bunlara ek olarak, biyoinformatik araçlar kullanılabilir.

Markır dizileri elde etmek için kullanılan kaynaklardan biri de İfade Edilmiş Ardışık Etiketlerdir (EST). EST'ler markır geliştirilmesi için özellikle çok çekicidirler, çünkü genomun protein kodlayan bölgelerinden ibarettir ve pek çok genom için çok hızlı adımlarla geliştirilmektedir. Bunun yanında son zamanlarda yapılan araştırmalarda incelenen birkaç bitki türünde EST'lerde genomik DNA ya göre mikro uydu frekansının daha yüksek olduğu gözlemlenmiştir. Bugüne kadar bitkilerde EST dizilerinden SSR'leri bulma işlemi tek çenekli bitkilere odaklanmıştır. İki çenekli bitkiler pamuk, soya fasulyesi, ayçiçeği, domates, patates vd gibi ekonomik öneme sahip bitkileri ve bitki genomics çalışmaları için model bitki Arabidopsis Thaliana'yı da içermektedir. Çift çenekli bitki türlerinin EST'lerinden SSR'lerin keşfi, farklı SSR kategorilerinde SSR'lerin bulunuş oranları ve frekanslarının bilinmesi sadece farklı türlerde SSR markırlarının geliştirilmesi için değil, aynı zamanda iki çenekli bitkilerde SSR'lerin fazla bulunmasının nedenini anlamak için de özellikle önemlidir. Bundan başka, EST'lerden elde edilen SSR'ler, esas olarak, ifade edilmiş gen dizilimleridir ve karşılaştırmalı genomik çalışmalarda potansiyel adaylardır.

2.4 Projenin Hedefleri

Ardışık tekrarları hesaplama yolu ile bulmada en önemli gereksinim etkin bir bilgisayar yazılımıdır. Her ne kadar günümüzde kullanımı kamuya açık birkaç yazılım varsa da bu yazılımlar bir veya birkaç yönden eksiktir. Bu yazılımlar uzunluğu sınırlı dizilerle çalışırlar ve bu dizi uzunlukları da genellikle 2 megabazı (Mb) geçmez. Çoğu yazılım da giderek önem kazanmakta olan tek nükleotitli ardışık tekarları dikkate almaz. Yazılımların çıktıları çok karmaşıktır. Bu yazılımları kullanan kişiler, sonuçları organize etmek ve yorumlamak için de çok zaman harcarlar. Bazı yazılımlar sonuçları doğrudan vermemekte, sonuçlar kullanıcının e-mail adresine gönderilmektedir. Nispeten etkin bazı yazılımlara da Web'ten ulaşmak, geliştirilmelerinden ve yayınlanmalarının üzerinden çok uzun bir süre geçmesine rağmen hala mümkün değildir. Örnek olarak, Valerio Parisi tarafından geliştirilen ve 2003 yılında yayınlanan STRING adlı yazılıma Web'ten ulaşmak hala mümkün değildir. Bu açıklamalardan

basit, etkin ve yüksek çıktılı (high throughput) ardışık tekrarları belirleme yazılımına ne denli çok gereksinim olduğu açıkca ortaya çıkmaktadır. Böyle bir yazılım ile bazı türlerde daha önce iç yüzü iyice açıklanamamış EST dizilerindeki SSR'ler de bulunacak veya bazı çalışmalar yapılmış olan türlerde de SSR'leri belirleme etkinliği artırılacaktır.

Bu tezin üç ana katkısı olacaktır:

- Etkin bir yazılımın geliştirilmesi ile büyük boyutlu veri dizilerinde de ardışık tekrarların bulunabilmesi
- Bu geliştirilmiş yazılım ile çok sayıda genom EST'lerinden SSR'leri bulma ve bu SSR'lerin frekanslarını ve dağılım yüzdelerini bulunabilmesi
- Markır geliştirilmesi ve diğer moleküler genetik analizler için çok önemli olan EST'lerdeki gereksiz fazla nükleotit dizilerinin belirlenmesi ve ayıklanması

3. YAZILIM ALTYAPISI

Ardışık Tekrarlı DNA dizilerini belirlemede kullanılan SSR'ler çok faydalı markırlardır. Fakat elde edilmeleri ve geliştirilmeleri çok zordur. Geliştirilmeleri için dizi hakkında ön bilgi gereklidir. SSR markırlarının geliştirilmesi genomik dizilerde ardışık tekrarların belirlenmesi ile başlar. Sonra SSR tekrar dizilerinin her iki ucuna bağlanan PCR primerlerinin dizayn edilmesi ile devam eder. SSR içeren dizilerin belirlenmesi için iki yaklaşım vardır.

1. Moleküler yaklaşım
2. Hesaplama yaklaşımı

Moleküler yaklaşımda önce SSR genomik kütüphaneleri oluşturulur, bunlar klonlanır ve elle veya bilgisayar yazılımları kullanılarak SSR motifleri belirlenir. Hesaplama veya biyoinformatik yaklaşımın moleküler yaklaşıma göre avantajı ise herkese açık veri bankalarından elde edilen dizilerin kolayca taranması ve bunların arasından SSR içerenlerin hızlı bir şekilde belirlenebilmesidir.

Kullanılan algoritmalara bağlı olarak SSR'lerin belirlenmesinde kullanılan hesaplama algoritmaları kabaca ikiye ayrılabilir:

1. Modele dayalı yaklaşımlar
2. Sözlük yaklaşımları

Modele dayalı yaklaşımlarda, ardışık tekrarlar için model tanımlanır ve bu model dizide tanıma uyan bölgeleri belirlemede kullanılır. Bu yaklaşım tam veya tam olmayan tekrarların çevresel bir listesini verir ve tekrar motiflerinin tanımlanması için motif tipi hakkında ön bilgiyi gerektirmez. Tandem Repeat Finder (Benson 1999) ve Sputnik

(Abajian 1994) gibi yazılımlar modele dayalı yaklaşımlardır. Tekrar eden dizi motifleri *a priori* (hipotez veya teoriye dayalı) olarak biliniyorsa, sözlük yaklaşımı motifleri belirlemede daha hızlı ve daha ölçeklendirilebilir çözümler sağlar. Bu yöntemde, yazılım verilen motiflerin sözlüğünü kullanır ve sözlüğün tüm içeriğini diziyi taramada kullanır.

3.1 İnformatik: DNA Dizilerinde Ardışık Tekrarların Yerini Belirlemede Kullanılan Kavram ve Algoritmalar

DNA dizisi, *S*, yazılım dilinde alfabesi $\sum = \{A,C,G,T\}$ olan *n* karakterli bir dizi olarak yorumlanır**.** Bilgisayar biliminde dizileri işleyen algoritmalar pek çoktur. Bu algoritmaların çoğu *S* altdizilerinde tam ve tam olmayan ardışık tekrarları bulmaya çalışır. *S* dizisi $S[i, j]$ olarak gösterilmektedir ve $1 < i < j < n$ notasyonu *i* pozisyonunda başlayan ve *j* pozisyonunda biten dizinin elemanlarını temsil eder.

3.1.1 Tam ve tam olmayan ardışık tekrarlar arasındaki benzerlik ölçümleri

Ardışık tekrarlar dizide en az iki defa tekrar eden alt dizilerdir. Ardışık tekrarlar denilince altdizilerin tam veya birebir tekrarı ve tam olmayan veya yaklaşık benzer tekrarları anlaşılır. Bu durum Tablo 3.1'de verilmiş olup dizinin motife üç değişik uyumu gösterilmiştir. Motif ve dizi arasındaki uyumsuzluklar koyu olarak gösterilmiştir.

Tablo 3.1 Motife tam ve benzer uyum durumları

Motif	**ACCGTGA**
Birebir (tam) uyum	ACCGTGA
3 adet uyumsuzluk gösteren benzer (tam olmayan) uyum. k=3 olan Hamming benzerlik ölçümü	AC**G**G**A**G**G**
1 silinme, 1 ekleme ve 1 adet uyumsuzluk olan benzerlik uyumu (biçimleme (edit) benzerliği) Levenshtein uzaklığı k=3	AA GTGGA

İki veya daha çok birebir uyum gösteren altdizinin ardışık tekrarları ifade ettiği fakat tam olmayan ardışık tekrarları belirleyebilmek için benzerlik ölçümlerine gerek duyulduğu görülmüştür. En tanınmış iki benzerlik ölçümü Hamming benzerlik ölçümü ve Levenshtein (edit) benzerlik ölçümüdür (Levenshtein 1966). Benzerlik ölçümleri bir diziyi diğerine dönüştüren bir seri işlem yaparak iki diziyi karşılaştırır. Hamming benzerlik ölçümü sadece uyumsuz eşleşmeleri bulur, edit benzerlik ölçümü ise uyumsuz eşleşmeleri, ayrıca nükleotit silinmesi veya eklenmesi durumundaki eşleşmeleri de bulur. Tek bir nükleotitin uyumsuzluğu $\sum$ alfabesinden bir karakterin başka bir karakter ile değişmesidir. Tek bir karakterin silinmesi durumunda diziden bir karakter çıkmıştır. Tek bir karakterin eklenmesi durumunda ise $\sum$ alfabesinden herhangi bir karakter diziye dahil olmuştur. Her bir benzerlik ölçümünün gerçekleştirilen işlem sayısına dayalı bir maliyeti vardır. Sonuçta elde edilen eşleşme en düşük maliyetle (en az işlem sayısı ile) bir diziyi diğerine dönüştüren bir işlem kümesidir.

Çoğu hesaplama algoritmaları, iki benzer dizi arasında izin verilen maksimum sayıdaki işlem sayısını veya maksimum maliyeti gösteren eşik değerlerini uygular. Örnek olarak, *k* sayıda yalnızca uyumsuzluk durumu olan problemlerde, bir algoritma en fazla *k* sayıda uyumsuzluğa sahip motifin dizideki tüm olasılıklarını bulur. Örneğin, *k* eşik değerli Hamming benzerlik ölçümünde uyumsuzluklar 1 maliyet değerine sahiptir. Diğer taraftan Levenshtein benzerlik ölçümünde *k* sayıda farklılık olan bir problem *k* sayıda uyumsuzluk, eklenme ve silinmeyi içerir. Örnek olarak *k* eşik değerli edit benzerlik ölçümünde her bir işlem bir maliyet değerine sahiptir.

3.1.2 Düzenli ifadeler (Regular expressions)

Düzenli ifadeler değişken sayıda karakter dizilerinden oluşan ancak belirli koşulları sağlayabilen ifadelerdir. Düzenli ifadeler yazılımdaki ihtiyaca göre düzenlenir. Diyelim ki bir metin dosyası içinde @ karakteri geçen bütün satırları elde etmek istiyoruz. Burada satırdaki karakterin uzunluğu ve ne olduğu önemli değil; yeter ki @ karakteri olsun. Belirtilen bu satırları elde etmenin çeşitli yolları olabilir. Ancak şartlarımız arttıkça işlemi koda dökmek zorlaşacaktır. Örneğin milyonlarca e-mail adresi olabilir. Ama bir tane e-mail adresi formatı vardır. Her e-mail adresi mutlaka @ karakteri ve en az bir '.' karakteri içermelidir. Eğer birden fazla nokta varsa, noktalardan biri mutlaka @ karakterinden sonra olmalıdır. Gördüğünüz gibi bir karakter dizisinin gerçek bir

e-mail adresi olup olmadığını test etmek bir hayli zor. Bu yüzden C#'ta bu tür düzenli ifadeleri temsil etmek için Regex sınıfı geliştirilmiştir. Regex sınıfı `System.Text.RegularExpressions` isim alanında bulunmaktadır. Bir karakter dizisinin, oluşturulan düzenli ifadeye uyup uymadığını belirlemek için ise yine aynı isim alanında bulunan `Match` adlı sınıftan faydalanılır.

3.1.2.1 Düzenli ifadelerin oluşturulması

1. Bir düzenli ifadenin satır başında mutlaka istenilen bir karakter ile başlanması isteniyorsa ^ karakteri kullanılır. Örneğin ^9 düzenli ifadesinin anlamı yazının mutlaka 9 ile başlaması demektir. "9Abcf" yazısı bu düzenli ifadeye uyarken "dasA" yazısı uymamaktadır.
2. Belirli karakter gruplarını içermesi istenen düzenli ifadeler için \ karakteri kullanılır. Örnek olarak; \D ifadesi ile yazının ilgili yerinde rakam olmayan tek bir karakterin bulunması gerektiği belirtilir. \d ifadesi ile yazının ilgili yerinde 0-9 arasında tek bir sayının olacağı belirtiliyor. \W ifadesi ile alfanümerik olmayan karakterin olması gerektiği bildiriliyor. Alfanümerik karakterler a-z, A-Z ve 0-9 aralıklarındaki karakterlerdir. \w ile yazıdaki ilgili yerde sadece alfanümerik bir karakterin olabileceği belirtilir. \S ifadesi ile yazının ilgili yerinde boşluk karakterleri (tab, space) dışında herhangi bir karakterin olabileceği bildiriliyor. \s ifadesi ile ilgili yerde sadece boşluk karakterlerinden birinin olabileceği bildirilir. Şu ana kadar gördüğümüz bilgiler ışığında ilk karakteri 5 ile başlayan ikinci karakteri herhangi bir sayı olan ve son karakteri de boşluk olmayan bir düzenli ifade aşağıdaki gibi gösterilebilir. Düzenli ifadeyi sağlayacak yazı mutlaka 3 karakterli olmalıdır. ^5\d\S ifadesinin tamamına filtre denilmektedir.
3. Belirtilen gruptaki karakterlerden bir ya da daha fazlasının olmasını istiyorsak + işaretini kullanırız. Örneğin; \w+ filtresi bir ya da daha fazla sayıda alfanümerik karakterin olabileceği anlamına gelmektedir. "2ASD" yazısı bu düzenli ifadeye uyarken "@Asc" yazısı uymaz.Çünkü @ karakteri alfanümerik değildir. + işareti yerine * işareti kullanırsak çarpıdan sonraki karakterlerin olup olmayacağı serbest bırakılır.
4. Birden fazla karakter grubundan bir ya da birkaçının ilgili yerde olabileceğini belirtmek istiyorsak mantıksal veya "|" operatörünü kullanırız. Örneğin; m|n|s düzenli ifadesi ile ilgili yerde sadece m, n ya da s karakterinin bulunabileceği

bildirilir. Bu ifadeyi parantez içine alıp sonunda + işareti koyarsak bu karakterlerden bir ya da birkaçının bulunabileceğini belirtmiş oluruz: (m|n|s)+

5. Sabit sayıda karakterin olmasını istiyorsak {adet} şeklinde belirtmeliyiz. Örneğin; \d{3}-\d{5} düzenli ifadesi ile "215-69857" yazısı sağlanır. Ama "54 56875" yazısı bu düzenli ifadeyi sağlamaz. Aradaki "-" işaretinin de mutlaka olması gerekir.
6. ? karakteri, kullanıldığı yerde önüne geldiği karakter en fazla bir en az sıfır defa olabileceğini bildirir. Örneğin; \d{3}B?A düzenli ifadesine "548A" ve "875BA" uyarken "478BBA" uymaz.
7. '.' işareti ile ilgili yerde '\n' karakteri dışında herhangi bir karakter bulunabilir. Örneğin; \d{3}.A düzenli ifadesine "587sA", "574AA", "8957A" yazıları uymaktadır.
8. \b ile bir kelimenin belirtilen karakter dizisi ile sonlanması gerektiği bildirilir. Örneğin; \d{3}dır\b düzenli ifadesine "584dır" ve "dsa325dır" yazıları uyarken "sda985dır8" yazısı uymaz.
9. \B ile bir kelimenin başında ya da sonunda olmaması gereken karakterler bildirilir. Örneğin; \d{3}dır\B düzenli ifadesine "584dır" ve "dsa325dır" yazıları uymazken, "sda985dır8" yazısı uyar.
10. Köşeli parantezler kullanarak bir karakter aralığı da belirtebiliriz. Örneğin ilgili yerde sadece büyük harf karakterlerinin olmasını istiyorsak [A-Z] şeklinde kullanmalıyız. Aynı şekilde küçük harf karakterleri için [a-z] kullanabiliriz. Bu aralık ilk ve son karakterler olmayabilir: Örneğin [A-P] ifadesi ile A ve P arasındaki karakterler alınır. Bu ifadeler sayılar için de geçerlidir. Örneğin [0-9] gibi.

Bu temel ifadeleri gördükten sonra C# programlama dilinde `Regex` ve `Match` sınıfları ile elimizdeki yazıların düzenli ifadelere uyup uymadığını nasıl bulacağımızı inceleyelim.

Regex sınıfı bir düzenli ifadeyi tutar.Bir Regex nesnesi oluşturmak için aşağıdaki kurucu metot kullanılabilir.

```
Regex  rgx = new Regex(string filtre)
```

filtre parametresi yukarıda anlatılan düzenli ifadeleri temsil etmek için kullanılan sembollerden oluşan bir yazıdır. `Regex` sınıfının `Match()` metodu kendisine gönderilen bir yazının düzenli ifadeye uyup uymadığını kontrol eder ve uyan sonuçları `Match` sınıf türünden bir nesne ile geri döndürür. `Match` sınıfının `NextMatch()` metodu ise verilen yazıda bulunan bir sonraki düzenli ifadeyi döndürür. Yazının düzenli ifadeye uyup uymadığının denetimi ise `Match` sınıfının `Success` özelliği ile yapılır. Eğer düzenli ifadeye uygun bir yapı varsa `Success` özelliği true olur.

C# programlama dilinde düzenli ifadelerle ilgili bir diğer sınıf ise `MatchCollection` sınıfıdır. Bu sınıf ile bir yazı içerisinde düzenli ifadeye uyan bütün `Match` nesneleri tutulur. `MatchCollection` nesnesi aşağıdaki gibi oluşturulabilir.

```
MatchCollection mc = Regex.Matches(str,filtre);
```

Burada `Regex` sınıfının statik `Matches( )` metodu kullanılmıştır. Bu metodun ilk parametresi kontrol edilmek istenen yazı, ikinci parametre ise düzenli ifadenin kendisidir. Bir `MatchCollection` nesnesi oluşturulduktan sonra foreach döngüsü yardımıyla bu koleksiyondaki bütün `Match` nesnelerine erişebiliriz. `Match` nesnesine eriştikten sonra düzenli ifadeye uyan karakter dizisinin orijinal yazıdaki yerini ve yazının kendisini `ToString()` metodunu kullanarak elde edebiliriz. `MatchCollection` sınıfının `Count` özelliği ile düzenli ifadeye uyan alt karakter dizilerinin sayısı verilir. Eğer `Count` özelliği sıfır ise düzenli ifadeye uyan yazı bulunamadı demektir.

Buraya kadar anlatılanlara bir örnek verelim. Düzenli ifademiz aşağıdaki gibi olsun;
A\d{3}(a|o)+

Bu düzenli ifade ile başlangıcı A karakteri olan ve bu karakterden sonra 3 tane rakam sonra da ‘a’ ya da ‘o’ karakterinden bir ya da birden fazla sayıda karakter grubu doğru kabul edilir.

3.1.3 Dinamik programlama ile dizileri hizalama

Dinamik programlama, diziler arasında optimal benzerliği bulmak için 2 veya daha çok dizinin bir başka diziye göre hizaya getirilmesinde kullanılan bir tekniktir. 1955 yılında Bellman matematiksel temellerini ortaya koyarak sistematik dinamik programlama çalışmalarına başladı (Bellman 1957). Biyoloji dalında önce çevresel (global) sonra da yerel (lokal) dizilimi çözmek için iki ana çalışma sunuldu. (Needleman ve Wunsch 1970) iki dizinin global dizilimi için çözüm sundular. (Smith ve Waterman 1981) yerel dizilim problemini çözdü. 1988 de, (Myers ve Miller 1988) sarmal dinamik programlamayı sundular. Fakat bu program (Fischetti vd 1992) sarmal dinamik programlamayı tekrar sununcaya kadar farkedilmeden kaldı. Konumuz ardışık tekrar paternleri ile DNA dizileri arasında benzerlik olduğu için sadece iki dizi arasındaki hizalama konusu ele alınacaktır. İlk olarak iki dizi arasındaki çevresel (global) hizalama, sonra iki dizinin alt dizileri arasındaki yerel (lokal) hizalama tanımlanacaktır.

3.1.3.1 Çevresel (Global) hizalama

Ele alınan S_1 ve S_2 gibi iki dizinin çevresel hizalanması, S_1 ve S_2 dizilerinin ya içine yada uçlarına boşluk yerleştirilerek elde edilir. Sonra boşluk içeren iki dizi öyle üst üste getirilir ki her iki dizide de boşluk karşısına karakter veya karakter karşısına boşluk gelir. Boşluklar bir dizide silinmeyi, karşı dizide ise eklenmeyi gösterir.

Örnek: ACGCTCTA ve ACCTATGA dizilerinin çevresel hizalanmasını inceleyelim.

AC**G**CT**C**T_A
AC_CT**A**T**G**A

Bu dizilimde, her iki G nin karşısında boşluk, koyu gösterilen C ve A arasında ise uyumsuzluk vardır. İki dizi arasındaki diğer tüm pozisyonlar uyum durumunu göstermektedir.

3.1.3.2 Yerel (Lokal) hizalama

Ele alınan S_1 ve S_2 gibi iki dizinin yerel hizalanması, S_1 dizisinde s_1 altdizisinin ve S_2 dizisinde s_2 altdizisinin ya içine yada uçlarına boşluk yerleştirilerek elde edilir. Sonra boşluk içeren iki alt dizi öyle üst üste getirilir ki her iki alt dizide de boşluk karşısına karakter veya karakter karşısına boşluk gelir. Boşluklar bir alt dizide silinmeyi, karşı alt dizide ise eklenmeyi gösterir.

Örnek: ACGCTCTA ve ACCTATGA dizilerinin altdizilerinde yerel hizalanmayı inceleyelim.

CT**C**T

CT**A**T

Bu hizalamada koyu yazılmış C ve A arasında uyumsuzluk vardır. Diğer tüm pozisyonlar iki alt dizi arasındaki uyumu gösterir.

3.1.3.3 Benzerlik ölçümü (Edit distance)

Yerel ve çevresel dizilim için yaptığımız uyumluluk, uyumsuzluk ve aralık gibi tanımlar benzerlik ölçümleri için de geçerlidir. Benzerlik ölçümü sabit iki dizinin hizalanmasındaki işlem sayısıdır. Dinamik programlama, iki dizinin bütün alt dizilerinin benzerlik ölçümlerinin hesaplanması için teknikler sağlar.

Benzerlik ölçümü; n karakter içeren S_1 çevresel $(1..i)$ dizisi ve m karakter içeren $S_2(1..j)$ dizileri için $T(i, j)$, s_1 ilk i karakterini s_2 nin ilk j karakterine dönüştürmek için gerekli minimum benzerlik ölçümü sayısını gösterir. Benzerlik ölçümü $T(n,m)$, S_1 ve S_2 dizilimlerinin çevresel (global) dizilimlerine karşılık gelir; $0 \leq i \leq n$ ve $0 \leq j \leq m$ aralığında değişen i ve j nin tüm kombinasyonları için benzerlik ölçümü $T(i, j)$ çözülür.

3.1.3.4 Dinamik programlama

Dinamik programlama iki dizinin hizalanmasını üç ana adımda gerçekleştirir:

i. Yineleme ilişkisi
ii. Matriste hesaplama
iii. Geri izleme yolu

Yineleme ilişkisi tüm kabul edilebilir benzerlik ölçümlerini tanımlar. T matrisi, yatay ve düşey eksenlerine hizalanacak diziler yerleştirilerek ve her bir hücre *T(i, j)* benzerlik ölçümünü temsil edecek şekilde oluşturulur. Geri izleme yolu iki diziyi hizalamada izlenecek benzerlik işlemlerinin sırasını açıkça belirler.

Yineleme ilişkisi, temel ve özyineleme (recursive) durumlarından oluşur. Herbir *T(i, j)* de, *i* ve *j* sırasıyla $0 \leq i \leq n$ ve $0 \leq j \leq m$ değerlerini aldığında skorlamanın nasıl hesaplanacağını tanımlar. Dizi veya alt dizinin başlangıcına veya sonuna boşluk yerleştirilmesinin skorlamaya hiçbir etkisi olmadığı için, tüm $0 \leq i \leq n$ ve $0 \leq j \leq m$ durumlarında; $T\ (i, 0) = 0$

$T\ (0, j) = 0$ dır.

Herbir *T(i, j)* hücresi matriste daha önce hesaplanmış değerler kullanılarak ve bu değerlerden uyum, uyumsuzluk ve boşluk durumlarına göre bu hücreye tekrar geçiş yapılarak hesaplanır. Dizilimde her boşluğa bir değer atandığı için, *T(i, j)*'yi hesaplamak için özyinelemeli koşul

$$T(i-1, j-1) + \text{uyum-testi}(i, j)$$
$$T(i-1, j) + \text{boşluk}$$
$$T(i, j-1) + \text{boşluk}$$

durumlarının en iyisidir. Hesaplamada i değerleri $1 \leq i \leq n$ ve $1 \leq j \leq m$ değerleri arasında değişir. Uyumluluk testinde S_1 [i] = S_2 [j] ise uyum-testi(i, j) uyumluluk durumunu; S_1 [i] $\neq$ S_2 [j] ise uyumsuzluk durumunu gösterir. Geleneksel benzerlik ölçümünde uyum, uyumsuzluk ve aralık için maliyet olarak 1 değeri atanır fakat farklı maliyetler kullanılması halinde alternatif optimizasyon durumları oluşur.

3.1.3.5 Sarmal dinamik programlama (Wraparound dynamic programming)

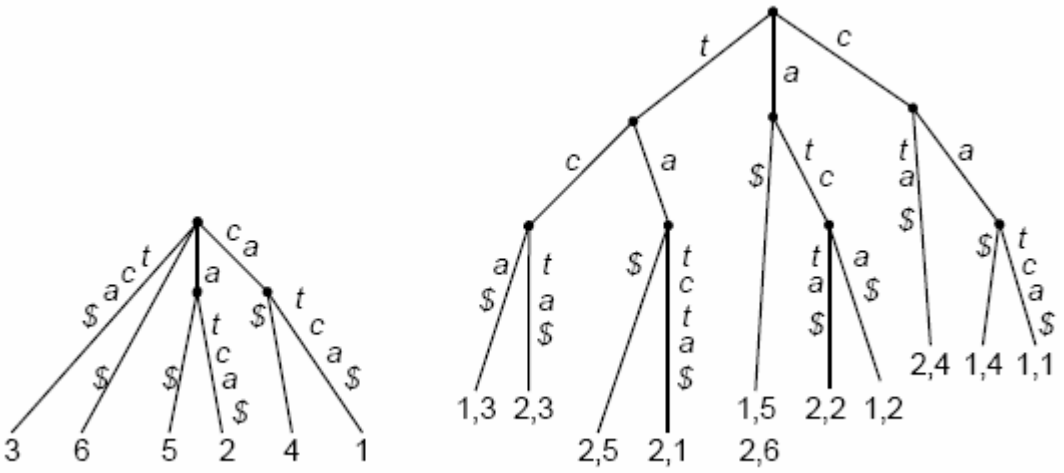
Sarmal dinamik programlama dizide ardışık tekrarların belirlenmesi için çok değerli bir tekniktir. Bu teknikte tüm dizi motifleri bilinen ardışık tekrar ile baştan aşağı taranır. Sarmal özelliği, standart dinamik programlama algoritmasında T(i, m)'den T(i, 1)'e ve T(i, m)' den T(i + 1,1)' e geçişi sağlayarak dinamik programlamanın kapsama alanını genişletir. Bu işlem matris oluşturma esnasında her bir matris hücresinden ikinci bir geçiş yapılarak yerine getirilir. Tablo 3.2'de Sarmal dinamik programlama algoritması verilmektedir.

Tablo 3.2 Sarmal dinamik programlama algoritması.

Her sıra için i = 1..n Geçiş 1: T (i , j) yi hesapla Her sütun için j = 1..m T (i , j) yi özyinelemeli ilişkiyi kullanarak hesapla T (i , m) yi T(i, 0) a kopyala Geçiş 2: T (i , j) yi güncelle T (i , j) den T (i , j -1) e geçiş yap T (i , j) yi özyinelemeli ilişkiyi kullanarak tekrar hesapla

3.1.4 Sonek ağaçları

Tam bir sonek ağacı dizideki tüm sonekleri temsil eder (Şekil 3.1). Sonek, dizi içinde herhangi bir karakterden başlayan ve dizi sonuna kadar devam eden bir altdizidir. Sonekler bir ağaca öyle şekilde yerleştirilirler ki birbirine benzer karakterler ile başlayan iki sonek, sonek ağacı boyunca aynı yolu izlerler. Yol kök düğümünden başlar ve sonekler arasında bir farklılık oluşuncaya kadar aşağı doğru ilerler. Farklılaşmanın başladığı noktadan itibaren soneklerin her biri ayrı yol izlerler.



Şekil 3.1 Tek bir dizinin (en solda) ve çift dizinin (en sağda) sonek ağaç gösterimleri.

Sonek, dizi içinde herhangi bir pozisyondan başlayan ve dizi sonuna kadar devam eden bir altdizidir. Soldaki ağaç catca dizisindeki sonekleri gösterir. Sağdaki ağaç

S_1=catca ve S_2=tatcta dizilerindeki tüm sonekleri içerir. Soneklerin okunması en üst kök düğümünden başlar ve yaprağa kadar devam eder. Soldaki ağaçta sayılarla gösterilen yapraklar dizide soneklerin başlama pozisyonlarını gösterir. Sağdaki ağaçta yapraklar iki sayı ile gösterilmiştir. İlk sayı dizi numarasını, ikinci sayı ise o dizideki sonekin başlangıç noktasını gösterir. Siyah noktacıklar ağaçtaki düğümleri gösterir ve iki veya daha fazla sonekin bölündüğü noktaları temsil eder. Düğümler arasındaki her bir bölüm sonekteki bir veya daha fazla karakteri gösterir. '$' simgesi dizi bitimini gösterir (Gusfield 1997).

Kavramsal olarak, ağaçtaki her düğüm düğüme giren tek bir dala sahiptir. Düğüme giren, tüm sonekler benzer karakter serilerine sahiptir. Her düğüm, düğümü terkeden bir veya birkaç dala sahiptir. Benzerliğin devam ettiği sonekler aynı dala geçerler, farklılığın olduğu sonekler farklı dallara geçerler.

3.1.5 Fourier metodu

DNA da tam olmayan ardışık tekrarları bulmak için bir başka ilginç yöntem de (Tran vd 2004) tarafından geliştirilmiştir. Bu Fourier analizlerine dayalı bir yöntemdir.

Bu yöntemle N boyutlu dizide bulunan tam ardışık tekrarları bulmak için izlenen yol aşağıda kısaca özetlenmiştir.

3.1.5.1 Adım 1: DNA dizisindeki 4 nükleotidi χ A [n], χ T [n], χ C [n], χ G [n] şeklinde alt dizilere dönüştürmek

α , $\sum$ { A, T, C, G} kümesinin bir elemanıdır. DNA dizisinin n. pozisyonunda bulunan α karakteri varsa χ_α değeri χ_α [n] = 1 değerini, aksi halde 0 değerini alacaktır. Bu yüzden, χ_α , DNA dizisinde α karakterinin olup olmadığını gösteren bir sinyal olacaktır. Örnek olarak, '**ACTGCTAGCAAT'** DNA dizisinin χ_α [n] bileşenleri gösterilmiştir. (Tablo 3.3)

Tablo 3.3 'ACTGCTAGCAAT' dizisinin χ α [n] bileşenleri

∑	**A**	**C**	**T**	**G**	**C**	**T**	**A**	**G**	**C**	**A**	**A**	**T**
χ A [n]	1	0	0	0	0	0	1	0	0	1	1	0
χ T [n]	0	0	1	0	0	1	0	0	0	0	0	1
χ C [n]	0	1	0	0	1	0	0	0	1	0	0	0
χ G [n]	0	0	0	1	0	0	0	1	0	0	0	0

3.1.5.2 Adım 2: Ortalamadan sapmaların fourier' e dönüşümü

Önce $m_\alpha = \frac{1}{N}\sum_{n=\alpha}^{N-1} x_\alpha[n]$ i bulalım ve

$0 \le f \le 0.5$ ve $\alpha \varepsilon \{A, T, G, C\}$ için *f* tepe değerleri

$$S_\alpha(f) = \frac{1}{N}\sum_{n=0}^{N-1}(x_\alpha[n] - m_\alpha)e^{-j2\pi fn}$$

formülü ile hesaplanır.

3.1.5.3 Adım 3: Fourier çarpım spektrumunun oluşturulması

Aşağıdaki formül ile çarpım spektrumu hesaplanır:

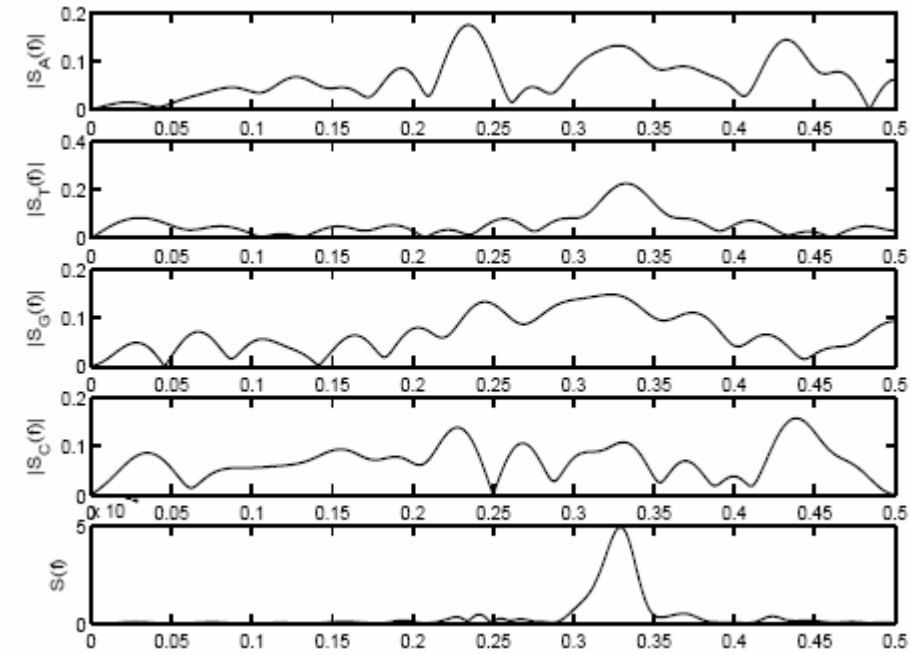
$$S(f) = \prod_{\alpha\in\{A,T,C,G\}}(|S_\alpha(f)| + c),$$

Burada *c* küçük pozitif bir sabittir. DNA dizisinde *P* tekrar periyodu varsa; *S(f)*, *f = 1/P* de bir tepe oluşturur. *S(f)* nin *f = 2/P, 3/P, ...* değerlerinde de tepe oluşturması mümkündür fakat sadece temel frekans gözönüne alınır. Böylece *P* periyodu tepe bölgesinden elde edilebilir. Özel bir karakter ardışık tekrarların bir parçası değilse, *c* sabiti *S(f)* nin boş çıkmasını önler.

3.1.5.4 Adım 4: Tam olmayan ardışık tekrar bölgelerinin başlangıç ve bitiş noktalarının belirlenmesi

Örnek: Aşağıda gösterildiği gibi gerçekte olmayan
ACTGACCGGACGC[ATGATGCTGATGATG]CTAC

gibi bir DNA dizisi alındığında, $|S_A(f)|$, $|S_T(f)|$, $|S_G(f)|$, $|S_C(f)|$ nın Fourier'e dönüşüm büyüklüklerini ve çarpım spektrumunu ($c = 0,01$ alındığında) göstermektedir. *S(f)* nin tepe noktası $f = 0,33$ de yerleşmiştir ve DNA dizisinde $P = 3$ ardışık tekrarı bulunur. Bu bilgiye dayanarak, motif TGA nın 14-28 pozisyonlarında 5 defa tekrar ettiği ve 20. pozisyonda 1 adet başka bir nükleotidle yer değiştirmenin gerçekleştiği belirlenmiştir. Şekil 3.2'de yukarıdan aşağıya doğru $|S_A(f)|$, $|S_T(f)|$, $|S_G(f)|$, $|S_C(f)|$, ve çarpım spektrumu *S(f)* gösterilmiştir.



Şekil 3.2 Tepe değerler ve Fourier çarpım spektrumu.

Tartışma bölümünde bu yöntem balarısı dizisi (GenBank: AMU73928) için diğer yöntemlerle karşılaştırılacaktır.

3.1.6 Gen bankası dizi koleksiyonu

Gen bankasından değişik özelliklere sahip bazı diziler seçilerek koleksiyon oluşturulmuştur. Bu koleksiyonu yapmanın ana amacı; geliştirilen algoritmanın geçerliliğini test etmek için referans noktaları oluşturmaktır. Dizi koleksiyonu basit ve karmaşık motif yapılarına sahip ardışık tekrar bölgelerini temsil eder. Bu koleksiyonda basit ardışık tekrarları (SSR), değişken uzunlukta ardışık tekrarlar (VLTR) ve değişken periyotlu birleşik ardışık tekrarlar gibi kısa ve uzun motifli bölgeler de vardır. Dizilerin

uzunluğu tek nükleotidli izole edilmiş ardışık tekrarları içeren kısa dizilerden birkaç yüz kilobaz çiftli orta uzunlukta diziler ve içinde pek çok ardışık tekrar bölgeler içeren çok uzun tam kromozom dizilerine kadar değişmektedir. Bundan başka, diziler bakteriden insana kadar geniş bir canlı yelpazesini içermektedir.

Bu diziler DNA içindeki ardışık tekrar bölgelerinde çok sık rastlanan genel problemler için de bir referans altyapısı oluşturacaktır. Böylece, bu diziler ardışık tekrarları belirlemek için yeni algoritma geliştirme çalışmalarında pek çok karmaşık ayrıntı içerdiği için çözülmesi gereken dizi parçaları olarak değerlendirilebilir. Tablo 3.4'de seçilen türlere ilişkin dizi uzunluklarıyla motif yapılarına yer verilmektedir. Tablo 3.5'de seçilen farklı dizilere ilişkin ardışık tekrar bölgelerinin içeriğine yer verilmektedir. Her dizinin ardışık tekrarlı bölgeleri takip eden Tablo 3.6'da verilmiştir. Tablo 3.6'da bölgeler dizi pozisyonları, motif yapı tipleri, motif dizileri veya motif uzunluğu ile tanımlanmıştır.

Tablo 3.4 Seçilen Gen Bankası dizi koleksiyonu. ▪ simgesi dizilerde dipnotlarına veya görsel çözümlemeye göre motif tiplerini göstermektedir.

Dizi	**Tür**	**Dizi Uzunluğu**	**Basit Motif Yapıları**			**Bileşik Motif Yapıları**	
Gen Bankası		(bp)	Genel	SSR	Uzun	VLTR	Birleşik
AMU73928	Balarısı	283		▪		▪	
BOVTGN	İnek	725		▪		▪	
BTA132392	İnek	251	▪	▪			
BTU75906	İnek	364		▪		▪	
DMPUGDMG1	Meyva Sineği	2468	▪	▪			
ECTRNYSU	Bakteri	1655	▪		▪	▪	
HSVDJSAT	İnsan	1985					▪
MM102B5	Fare	704					▪
MMMSAT5	Fare	412		▪			
U00144	İnek	407		▪			

Tablo 3.5 Gen Bankası koleksiyonundan seçilen farklı dizilerin ardışık tekrarlı bölgelerinin içeriğinin özeti.

Gen Bankası Lokusları	**Yeri**	**Dizideki Bilgiler**	**Görsel Gözlem**
AMU73928	76...209	Miniuydu tekrarlar	17 bp uzunluğunda SSR olmayan VLTR ve tek (T) motifli SSR
BOVTGN	311...703	46 ilâ 82 bp uzunluğunda 7 kopyaya sahip miniuydu tekrarları	23-28 bp uzunluğunda SSR olmayan VLTR ve çift (TG) motifli SSR
BTA1323392	69...242	24 ilâ 27 motif uzunluğunda 7 ardışık tekrarlı prion proteini	24 bp motifli ardışık tekrarlar
BTU75906	1...364	48 ile 79 bp uzunluğunda 5 kopyalı miniuydu tekrarları	23-28 bp uzunluğunda SSR olmayan VLTR ve çift (GT) motifli SSR
DMPUGDDMG1	2405...2468	TCTCTCT motifine sahip ardışık tekrar içeren göz pigment enzimi	1) TCTCTCT motifin 25 tam kopyasını içeren büyük bölge, pek çok benzer kopyalar 2) Yanlarında CT iki SSR içeren büyük bölge
ECTRNYSU	625...1158	Üç kopyalı 178 bp uzunluğunda motif	1) Üç kopyalı 178 bp uzunluğunda motif içeren ardışık tekrar 2) ACC motifine sahip yuvalanmış SSR
HSVDJSAT	1200...1543	11 kopyalı motif oluşturan yakın ilişkili 9 ve 10 bp uzunluğunda motiflerin 36 kopyası	CTGGGAGAGG, CTGGGAGAG ve CTGGGATTG üçlü motifine sahip karmaşık birleşik motif
MM102B5	1...696	234 bp uzunluğundaki tekrar	58 bp uzunluğunda ana motifli birleşik pattern
MMMSAT5	23...213	AC, AT ve GT motiflerine sahip SSR karışımını içeren mikro uydu bölgesi	AC, AT ve GT motiflerine sahip SSR demeti
U00144	292...407	AG, GT, ve ACAG motiflerini içeren SSR demeti	AG, GT, ACAG, AGGG ve CCGGGG motiflerine sahip birkaç SSR

Tablo 3.6 Gen Bankası koleksiyonundan alınan dizilerin seçilme nedenleri

GenBank Lokus: AMU73928. Bu dizi değişken kopya sayılı T motifi içeren çok iyi korunmuş VLTR içerir. SSR olmayan motif 17 bp uzunluğundadır. Bu bölge tek nükleotidli SSR içeren VLTR' ye iyi bir örnektir.
GenBank Lokus: BOVTGN. Bu dizi değişken kopya sayılı GT motifi içeren çok iyi korunmuş VLTR bölgelerine sahiptir. SSR'ler , SSR olmayan 23-28 bp lik motifler içine yuvalanmıştır. Bu bölge, iyi korunmuş, her bir kopyası kolaylıkla ayırt edilebilen SSR'leri içeren VLTR ye iyi bir örnektir ve yuvalanmış SSR'ler çok değişkenli kopya sayılarına sahiptir. Böylece, tam bölgeyi belirlemenin tek yolu VLTR'leri belirlemekten geçer. Ayrıca, bu bölge iki farklı özelliği de içerir. Bu özelliklerden birincisi, SSR ile SSR olmayan bölgeler arasındaki eklem yerinde, kopyalardan biri 7 adet G nükleotidi, diğer 4 kopya ise aynı eklem yerinde 3 adet G nükleotidi içerir. İkincisi, her SSR'nin öncesinde bulunan TGG nin ve sonrasında bulunan TG nin SSR ye mi yoksa SSR olmayan bölgeye mi dahil edileceğidir. G nükleotidlerin bulunması ve SSR'lerin yakınlarında farklı nükleotidlerin bulunması SSR olmayan dizilerin belirlenmesini güçleştirir.
GenBank Lokus: BTA132392. Bu dizi 24 ilâ 27 bp motifli yaklaşık yedi kopyalı ardışık tekrarlı bölgeleri içerir. Bu bölge hem yuvalanmış ardışık tekrarların hem de yüksek sıralı periodisite (aralıklı tekrar) gösteren bileşik tekrarların ipuçlarını taşır. Basit ve bileşik motif içeren dizi sınırında bulunduğu için bu dizi örnek olarak seçilmiştir.
GenBank Lokus: BTU75906. Bu dizi değişken kopya sayılı GT – motifler içeren çok iyi korunmuş VLTR bölgeleri içerir. Yuvalanmış SSR ler 23-28 bp lik SSR olmayan motiflerle birleşmiştir. Bu bölge kolayca tanımlanabilen yuvalanmış SSR içeren VLTR'lere çok iyi bir örnektir. Bu dizinin seçilme nedeni SSR'ler de bulunan düzensizliklerdir. Bu düzensizlikler benzer motife sahip olmalarına rağmen tek bir SSR yerine 2 SSR'nin tanımlanmasına yol açabilir. Bu tür düzensizlikler VLTR analizlerini daha da güçleştirmektedir.
GenBank Lokus: DMPUGDMG1. Bu dizi basit TCTCTCT motifini içeren bir dizidir. Bu motif ardışık 25 kopya boyunca çok iyi korunmuştur. Bu dizinin seçilme nedeni TCTCTCT motifi, CT motifi ve CT motifine sahip yuvalanmış SSR içeren VLTR bölgelerinin birbirinden ayırt edilebilmesidir.

GenBank Lokus: ECTRNYSU. Bu dizi çok iyi korunmuş 178 motifli üç kopyalı ardışık tekrarlı bölge içerir. Bu dizi büyük boyutlu motif içerdiği için seçilmiştir. Ayrıca bu dizinin yuvalanmış olarak ACC motifini içeren bir SSR sahip olduğu bulunmuştur.
GenBank Lokus: HSVDJSAT. Bu dizi üç temel motifi (CTGGGAGAGG, CTGGGAGAG ve CTTGGGATTG) içeren çözümlenmesi zor bileşik motifli bir bölgedir. Birisi 10 bp uzunluğunda iken diğerleri 9 bp uzunluğundadır. Basit seviyede, 10 motifli dizi 9 motifli dizi ile ardışık tekrarlanır. 9 bp motifin ardışıklığı , her dördüncü kopyada aynı motife tekrar rastlanılır. Bazen 10 bp lik dizi 2defa ardı ardına gelebilir. Bu durum her 11 lik döngüde bir defa oluşmaktadır.
GenBank Lokus: MM102B5. Bu dizi insan X kromozomundaki 234 bp uzunluğundaki gamma uydu bölgesinden bir parçadır. Gözle yapılan analizde bileşik bölge 58 bp lik temel bir baz serisine sahiptir. Ek motiflerle iki kopya 116 bp motife ve 4 kopya 231 bp motife dönüşür. 231 lik pb motifi çok iyi korunmuştur ve 234 bp lik standart gamma uydu motifine yaklaşır. Çok iyi korunmuş birleşik motiflere örnek olarak seçilmiştir.
GenBank Lokus: MMMSAT5. Bu dizi AC, AT ve GT motiflerini içeren SSR demeti içerir. Bu bölgedeki hemen hemen her pozisyon bu üç motife dayalı kabul edilebilir. Motif karışımından oluştuğu için bu dizi seçilmiştir.
GenBank Lokus: U00144. Bu dizi AT, GT, ACAG, AGGG ve CCGGG motiflerini içeren SSR demetine sahiptir. Aynı motif uzunluğunu sahip birkaç SSR nin belirlenmesi için seçilmiştir.

3.2 SSR'leri Bulan Yazılımların İncelenmesi

Bu çalışmada, dizideki hemen hemen tüm SSR tiplerini bulabilen etkin, basit ve yüksek çıktılı bir yazılım geliştirilmesi hedef alınmıştır. SSR'leri belirleme yazılımının geliştirilmesine başlanmadan önce mevcut olan yazılımların SSR li dizileri bulmada aşağıda belirlenen ölçütlere ne oranda yaklaştığı araştırması yapıldı. Aşağıdaki ölçütler arzu edilen bir SSR belirleme yazılımında bulunması gereken asgari ideal özellikler olarak ele alındı:

- Tekli, ikili, üçlü ve dörtlü nükleotid tekrarlarını belirleyebilme yeteneği,

- Birleşik tekrarları bulabilme yeteneği, örnek olarak iki veya daha fazla birleşik tekrar motiflerini bulabilme,
- Çok sayıda farklı SSR tiplerini belirleyebilme ve verilen dizide başlangıç ve bitiş yerlerini belirleyebilme yeteneği,
- Araya giren işe yaramayan diziler tarafından durdurulmuş kesikli tekrarları belirleyebilme,
- Çok sayıda diziyi içeren veri yığınlarında SSR'leri belirleme işlemini yapabilme,
- Çıktıda en azından mutlaka dizi ismi, motif tipi ve ardışık tekrarların başlangıç ve bitiş yerlerinin verilmiş olması,
- Yazılımın Web te sitesi olup olmadığı veya Web ten yazılıma ulaşılabilme olasılığı

Bu ölçütlere göre incelenen bazı yazılımlar aşağıda verilmiştir:

3.2.1 Sputnik

Bu yazılım C programlama dilinde yazılmış, FASTA formatındaki DNA dizilerinde mikro uydu tekrarlarını bulan basit bir yazılımdır (Abajian 1994). Dizi dosyası yazılıma girdi olarak verilir ve standart çıktıda, tekrarların dizideki yeri, uzunluğu ve hata sayısı alınır. Sputnik 2 ile 5 nükleotid uzunluğundaki tekrar motiflerini bulmak için geliştirilmiştir. Eklemeler, uyumsuzluklar ve silinmeler olması gerekenin çok altında bulunur. Sputnik düşük çıktılı uygulamalar için uygundur. Tek nükleotidli ardışık tekrarları bulamaz. Web desteği de yoktur.

3.2.2 FindPatterns

Bu yazılım, şimdi adı Accelry olan Genetics Computer Group (GCG)'un biyoinformatik yazılımlarından biridir. Bu yazılım www.accelrys.com adresli siteden temin edilebilir. Büyük veri kümelerini baştan sona inceler, kullanıcı tarafından tanımlanan kısa nükleotid ve amino asit motiflerini belirler. FindPatterns uyumsuzluk içeren bazı motifleri fark edebilir fakat kesikli motifleri bulamaz. FindPatterns sonuçları çıktı dosyasında verir ve bu dosya doğrudan diğer yazılımlarda da kullanılabilir. Bununla birlikte, FindPatterns kesikli motifleri bulmada ki eksiklikleri yüzünden, birleşik tekrarları etkili bir şekilde bulamaz.

3.2.3 Repeat Finder

Repeat Finder, özellikle SSR'leri belirlemek için geliştirilmiş web-içerikli bir yazılımdır (WEB_2 2006). Başlangıçta bu yazılım tek girdili dizilerdeki SSR'leri bulmak için geliştirilmişse de sonraları yazılıma çoklu veri yığınlarını da işleyebilme özelliği kazandırılmıştır. Küçük ve orta boyutlu veri kümelerinde, RepeatFinder iyi bir yazılım olmasına rağmen, son sürümünde aşağıda belirtilen kısıtlamalara sahiptir:

1. Tek nükleotidli tekrarları bulmaz
2. Çoklu veri yığınlarında düşük performans gösterir
3. 3 Kb (Kilobaz) dan büyük dizilerde yazılımın hızı önemli derecede düşer;
4. Çıktı basit, uzun bitişik zincirleme şeklindedir ve tek tek ardışık tekrar dizilerinin açığa çıkartılması büyük zaman kaybına neden olur.

3.2.4 Tandem Repeats Finder (TRF)

TRF; ardışık tekrarları gösteren ve yerlerini belirleyen bir yazılımdır (Benson 1999). Yazılımı kullanmak için, kullanıcı verilerini FASTA formatında vermelidir. Motifin tanımlanması, uzunluğunun bildirilmesi ve diğer parametrelerinin hiçbirinin belirtilmesine gerek yoktur. Çıktı 2 dosya içerir: Ardışık Tekrar çizelgesi ve dizilim dosyası. Tekrar çizelgesi, bulunan herbir tekrar hakkında bilgiler içerir. Bu bilgilere ardışık tekrarın yeri, uzunluğu, kopya sayısı ve nükleotid içeriği dahildir. Çizelgedeki yer indekslerinin üzerine tıklanması ile ikinci web tarayıcı sayfası açılır ve uzlaşılan motife göre kopyaların hizalanmasını gösterir. Yazılım çok hızlıdır ve birkaç saniye içinde 5 Mb (Megabaz) uzunluğundaki dizileri analiz eder. Diziler istenilen boyutta olabilir. Ardışık tekrarların motif boyutları 1 ile 2000 baz arasında olabilir. Sunucuya gönderilen dizi bilgileri güvence altındadır ve yazılımın yürütülmesinden sonra silinir. Görüldüğü gibi, TRF belirlediğimiz tüm ölçütlerin tamamını karşılamaktadır.

3.3 Tandem Repeats Miner Yazılımının Geliştirilmesi

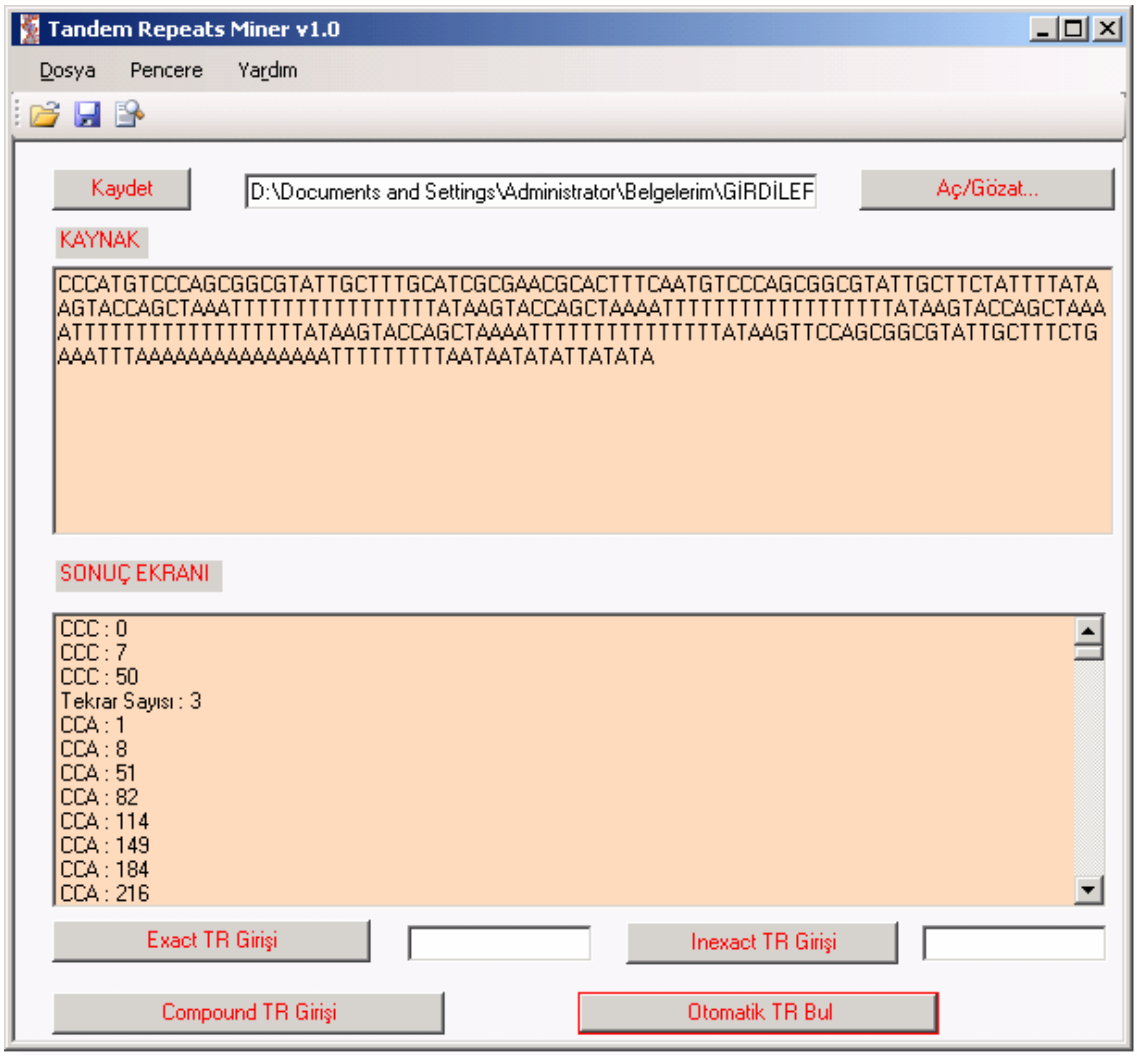
Görüldüğü gibi, incelenen SSR belirleme yazılımları böyle bir projeden istenen hızlı, etkin ve ayrıntılı çıktı gibi özellikleri sağlamada bir veya birkaç yönden yetersiz

kalmaktadır. Bu yüzden arzu edilen ölçütleri karşılayacak yeni bir yazılım geliştirilmesine başlanmıştır. Bu yazılıma Tandem Repeats Miner adı verilmiştir. Uygulama Visual C# dili kullanan nesneye dayalı bir yazılımdır. Kullanıcıya uygun etkileşimli bir arayüzü vardır ve Windows platformunda çalışır. DNA veri bankalarından alınan Fasta formatlı dizilerde ve EST'lerde SSR'leri bulur. Tandem Repeats Miner yazılımı aşağıda kısaca özetlenmiştir:

- Yazılım Visual C# dilinin motif tanıma (düzenli ifade) kavramı kullanılarak geliştirilmiştir.
- Yazılımda tek, çift, üç ve dört nükleotidli ardışık tekrarların minimum sayısı ve ardışık tekrarlar arasındaki artık dizilerin dikkate alınacak uzunluğu isteğe bağlı olarak ayarlanabilir.
- Dizi dosyalarını FASTA formatında kabul eder. ">" işaretinden ve dizi isminden sonra ilk nükleotidden son nükleotide kadar diziyi tarar, boşlukları ve FASTA'ya ait özel terimleri siler ve her bir sıra işlendikten sonra \n özelliğini de elemektedir.
- (2-4) uzunluğunda veya (1-4) uzunluğunda motifli tekrarların bulunması işlemini gerçekleştirir. Dizi uzunluğunu bulur, dizi uzunluğunu ardışık tekrar sayısına böler ve tekrarın uzunluğunu bulur. Ardışık tekrar uzunluğunu tüm dizi boyunca kaydırarak diziyi tarar. Bu işlemi diğer tüm ardışık tekrar motifleri için de tekrarlar. Dizi tamamlandığında ve yeni ">" işaretine rastlandığında dosyanın diğer tüm dizileri için de aynı işlemleri tekrarlar.
- Dosya veya dizi içindeki dizi adlarını, ardışık tekrar tiplerini, tekrar sayılarını ve tekrarların başlangıç ve bitiş noktalarını kaydeder.

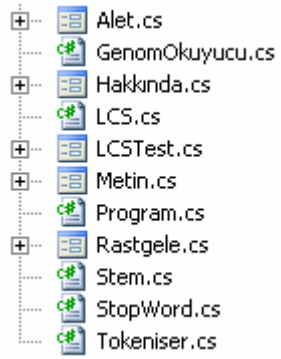
3.4 Tandem Repeats Miner Yazılımının Arayüzü

Uygulama geliştirilirken kodlama için Microsoft® Visual Studio 2005 kullanılmıştır. Şekil 3.3'de Tandem Repeats Miner yazılımının kullanıcı grafik ara yüzüne ait ekran görüntüsü verilmiştir.



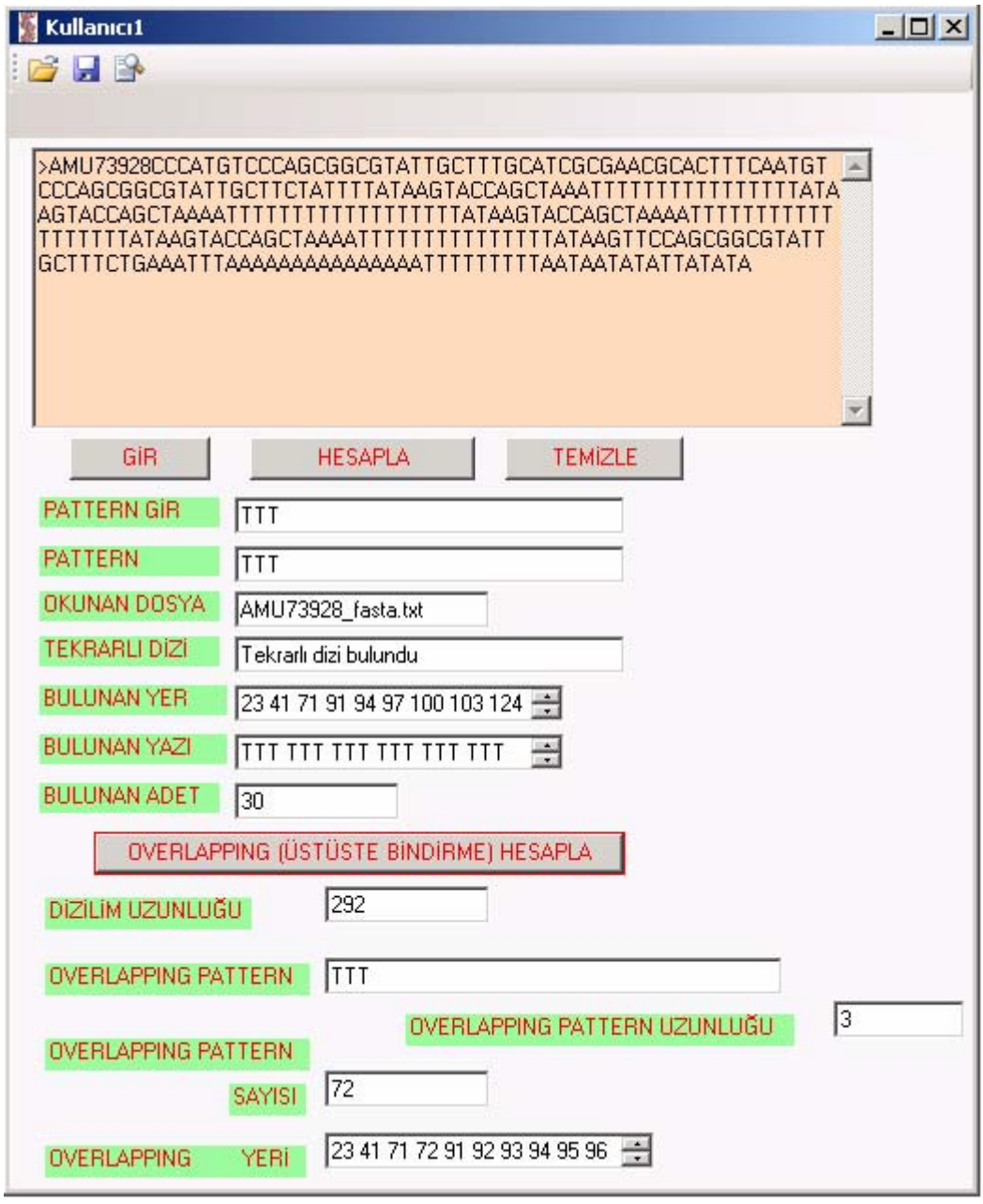
Şekil 3.3 Tandem Repeats Miner grafik ara yüzü

Geliştirilen yazılım bir MDI (Multiple Document Interface – Çoklu Belge Arayüzü) olup form ve sınıf dosyaları Rastgele adı verilen ana bir form üzerinden çalıştırılmaktadır. Geliştirilen form ve sınıf dosyaları Şekil 3.4'de verilmektedir.



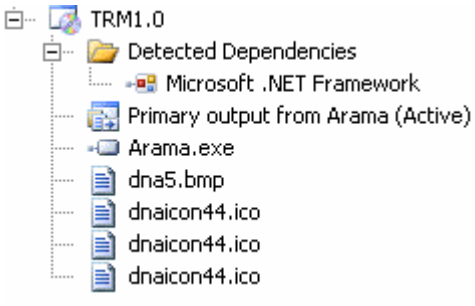
Şekil 3.4 Tandem Repeats Miner sınıf yapısı

Tandem Repeats Miner aynı zamanda kullanıcı girişli bir program olup overlapping (üst üste binme) şeklinde sonuçları bulması hususunda kullanıcıya ayrı bir seçenek sunmaktadır. Şekil 3.7'de ilgili forma ilişkin ekran görüntüsü verilmektedir.



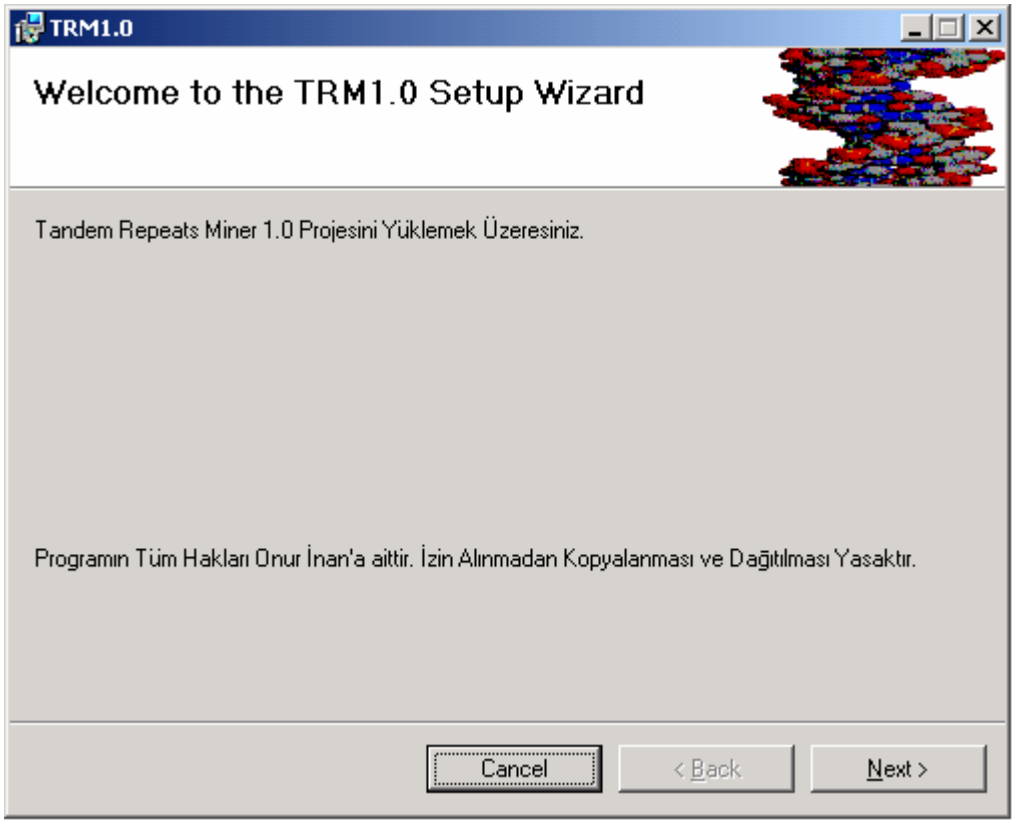
Şekil 3.5 Tandem Repeats Miner kullanıcı grafik ara yüzü

Yazılım; ilgili konuda çalışma yapmak isteyen araştırıcıların hizmetine sunulması amacıyla bir paket program haline getirilmiştir. Bu sayede yazılımın setup dosyaları internet ortamına aktarılarak serbestçe kullanıma sunulacaktır. Geliştirilen setup dosyaları Şekil 3.6'da verilmektedir.



Şekil 3.6 Tandem Repeats Miner setup dosyaları

Yazılım Windows işletim sistemi altında .NET Framework yazılımının 2.0.50727 ve üstü versiyonuna sahip olunması halinde çalıştırılmaktadır. Şekil 3.6'da program kurulumuna ilişkin ekran görüntüsü yer almaktadır.



Şekil 3.7 Tandem Repeats Miner kurulumu

4. BULGULAR VE TARTIŞMA

Bu bölümde gen bankası dizileri yazılım algoritması kullanılarak analiz edilmekte ve sonuçları mevcut yazılım sonuçlarıyla karşılaştırılarak değerlendirmesi yapılmaktadır. Bir önceki bölümde anlatılan değişik özelliklere sahip diziler yazılıma girilerek algoritmanın geçerliliği test edilmektedir. Aşağıda yer alan diziler gen bankası dizi koleksiyonuna dahil olup bakteriden insana çeşitli canlılardan örnekler içermekte ve değişik motif yapılarına sahip ardışık tekrar bölgelerini temsil etmektedir.

4.1 GenBank Lokus – AMU73928

Bu dizinin alındığı tür *Apis mellifera* (balarısı) dır.

4.1.1 Genbank dizi bilgileri

Dizi dipnotlarında miniuydu bölgelerinin 76'ıncı pozisyondan başlayarak 209'uncu pozisyona kadar devam ettiği belirtilmiş fakat ardışık tekrarlı bölgeler için motifler belirtilmemiştir.

4.1.2 Görsel analiz

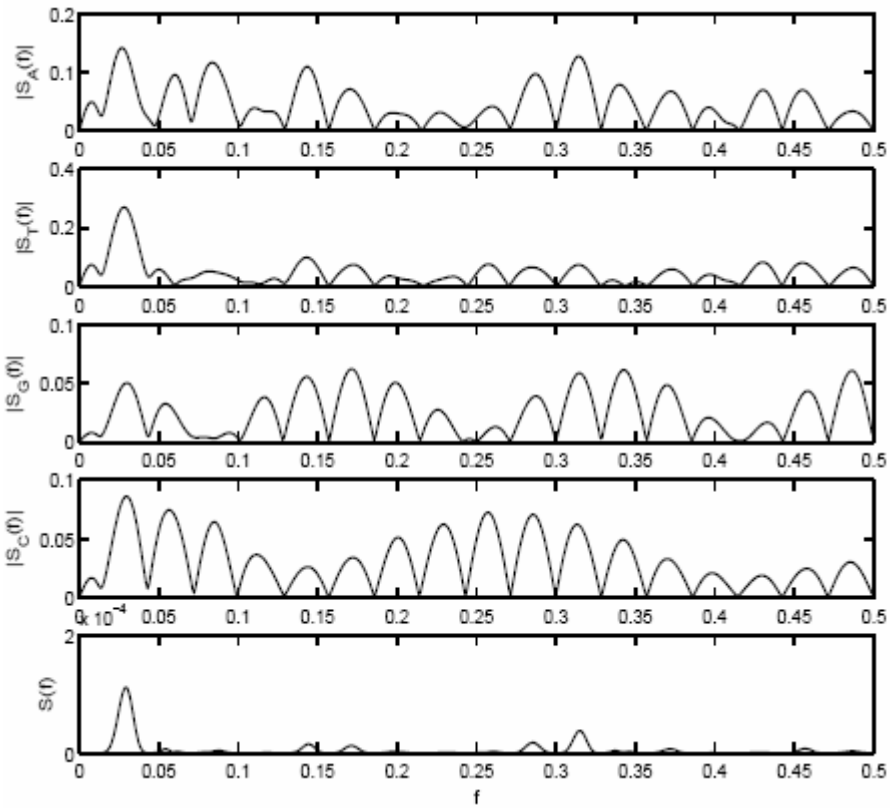
Dizi iyi korunmuş; 17 bp uzunluğunda SSR olmayan değişken uzunlukta ardışık tekrarları (VLTR) ve VLTR bölgesi içinde yuvalanmış T motifine sahip SSR'leri içermektedir.

4.1.3 Algoritmanın performansı ve diğer yazılımlarla karşılaştırılması

Tandem Repeats Miner 72. ve 209. indisler arasında 4 tane ard arda 35 dizilim

uzunluğunda ardışık tekrar bulmaktadır. Bunlardan ilkinde 3 adet InDel (Insertion – Deletion) yer almaktadır. Diğerleri tam ardışık tekrarlardır. Consensus motifle (mutabakata varılmış ortak motif) birebir uyum göstermektedir. Tandem Repeats Finder (TRF) (Benson, 1999) yazılımı ve (Hauth ve Joseph, 2002) yine aynı sonucu tespit etmişlerdir. Consensus motif;
TTTTATAAGTACCAGCTAAAATTTTTTTTTTTTT şeklindedir.
17 bp uzunluğundaki VLTR'nin tespit edildiği gözlenmiştir.

(Tran vd 2004) sinyal işleme tekniğini kullanarak 65. ve 256. indisler arasında 5 tane 35 dizilim uzunluğunda ardışık tekrar bulmaktadır. Bulunan motif yukardaki sonuçlarla tutarlıdır. Ancak çok sayıda InDel yer almaktadır. Dizilim uzunluğunun tespiti Şekil 4.1'deki Fourier çarpım spektrumuna bakılarak anlaşılabilir. Yukarıdan aşağıya doğru $|S_A(f)|$, $|S_T(f)|$, $|S_G(f)|$, $|S_C(f)|$, ve çarpım spektrumu $|S(f)|$ gösterilmiştir.



Şekil 4.1 AMU73928 için tepe değerler ve Fourier çarpım spektrumu.

S(f) nin tepe noktası *f = 0,0285* de yerleşmiştir ve DNA dizisinde *P = 35* ardışık tekrarı bulunur. Sinyal işleme tekniği daha çok büyük dizilim uzunluğuna sahip motifleri ya da az sayıda tekrar eden motifleri tespit etmek amacıyla kullanılır.

Tandem Repeats Finder (TRF) (Benson, 1999) yazılımı yukarıdaki motife ek olarak 72. ve 221. indisler arasında 2,2 tane ard arda 67 dizilim uzunluğunda ardışık tekrar tespit edilmektedir. Toplam 6 adet InDel yer almaktadır. 65 dizilim uzunluğundaki consensus motif;
TTTTATAAGTACCAGCTAAATTTTTTTTTTTTTTTATAAGTACCAGCTAAAA
TTTTTTTTTTTT şeklindedir. Bu motif Tandem Repeats Miner da kullanıcı girişli olarak tespit edilebilmektedir.

4.2 GenBank Lokus – BOVTGN

Bu dizinin alındığı tür *Bos Taurus* (inek) dur.

4.2.1 Genbank dizi bilgileri

Bu dizi BTGL1 ile gösterilen içinde SSR'lerin yerleştiği değişken kopya sayılı miniuydu bölgesini içerir. BTGL içeren parça uzunluğu değişkenlik gösterir ve parmak izi saptamasında kullanılır (Nave vd 1997). Bu dizi 46 ve 82 bp uzunluğunda 7 kopyalı miniuydu ve her bir kopya için GT motifli SSR'lerle birleşmiş 29 bp uzunluğunda bir dizi içerir (Nave vd 1997).

4.2.2 Görsel analiz

Görsel analizle, GenBank da belirlenen özellikler tespit edilmiştir. Motif bileşik motif yapısındadır ve içinde bulunduğu bölge değişken uzunlukta ardışık tekrar bölgesi olarak sınıflandırılmıştır. Dizide tam SSR dizisinden sonra iki ardışık T nükleotid içeren TG dizisi belirlenmiştir.

4.2.3 Algoritmanın performansı ve diğer yazılımlarla karşılaştırılması

Tandem Repeats Miner yazılımı 327. ve 352. indisler arasında 13 tane ard arda 2 dizilim uzunluğunda ardışık tekrar bulmaktadır. Motif GT şeklindedir. Aynı motifi 425. ve 462. indisler arasında 19 kere; 641. ve 690. indisler arasında 25 kere bulmaktadır. Bütün tekrarlar tam ardışık tekrardır. Consensus motifle birebir uyum göstermektedir.

TRF (Benson, 1999) yazılımı ve (Hauth ve Joseph, 2002) yine aynı sonucu tespit etmişlerdir. (Hauth ve Joseph, 2002) aynı SSR'leri 23-28 bp'lik motifler içerisinde serpiştirilmiş şekilde bulmuştur.

Tandem Repeats Miner yazılımı 330. ve 425. indisler arasında 2 tane ard arda 48 dizilim uzunluğunda ardışık tekrar bulmaktadır. İkinci tekrarda bir adet eşleşmeme durumu vardır. Consensus motif;
TGTGTGTGTGTGTGTGTGTGTTGCCTGTCTCCAGCGTAAGTAATCA
şeklindedir. Tandem Repeats Finder (TRF) (Benson, 1999) yazılımı aynı consensus motifi 330. ve 448. indisler arasında 2,5 kere bulmaktadır.

Tandem Repeats Miner yazılımı 378. ve 501. indisler arasında 2 tane ard arda 62 dizilim uzunluğunda ardışık tekrar bulmaktadır. Birinci tekrarda üç adet eşleşmeme durumu vardır. Consensus motif;
TGTGTGTGTGTGTGTGTGTGTTGCCTGTCTCCAGAGTAAGTAATCATGGG
TGTGTGTGTG şeklindedir. Tandem Repeats Finder (TRF) (Benson, 1999) yazılımı aynı consensus motifi 378. ve 508. indisler arasında 2,1 kere bulmaktadır.

4.3 GenBank Lokus – BTA132392

Bu dizinin alındığı tür *Bos Taurus* (inek) dur.

4.3.1 Genbank dizi bilgileri

Bu dizi prion proteini geninden alınmış bir bölgeyi temsil eder. 24 ile 27 bp uzunluğunda 7 kopyalı bir bölgeyi içerir. Uzunluk farklılığı, 24 veya 27 nükleotid motifli bölge içinde yerleştirilmiş GGT motifine sahip tam olmayan SSR lerden kaynaklanır.

4.3.2 Görsel analiz

Dizi hakkında literatür bilgileri görsel olarak doğrulanmıştır. Yerleştirilmiş SSR'lerin analizi ile bölgenin başlangıç ve sonunda bulunan motiflerin 27 nükleotidli

olduğu ve motiflerin tam olmadığı görülmüştür. Bu bölgedeki SSR'lerin iyi korunmadığı halbuki değişken uzunluklu ardışık tekrarların iyi korunduğu saptanmıştır.

4.3.3 Algoritmanın performansı ve diğer yazılımlarla karşılaştırılması

Tandem Repeats Miner yazılımı 81. ve 224. indisler arasında 6 tane ard arda 24 dizilim uzunluğunda ardışık tekrar bulmaktadır.. İlk dört tekrarda toplam 9 adet eşleşmeme durumu vardır. Son iki tekrar tam ardışık tekrardır. Bulunan consensus motif GGTGGCTGGGGACAGCCACATGGT şeklindedir. Tandem Repeats Finder (TRF) (Benson, 1999) yazılımı aynı consensus motifi 81. ve 229. indisler arasında 6,2 kere bulmaktadır. (Hauth ve Joseph, 2002) aynı tekrarı ard arda 7 defa tespit etmektedir.

4.4 GenBank Lokus – BTU75906

Bu dizinin alındığı tür *Bos Taurus* (inek) dur.

4.4.1 Genbank dizi bilgileri

Bu dizi yukarıda bahsedilen BTGL miniuydu bölgesi için diğer alel (bir genin değişik formları) dizilimini içerir. Bu alel esas olarak BOVTGN de olduğu gibi 29 bp uzunlukta SSR olmayan motif ve bu bölge içinde yerleştirilmiş SSR içermektedir. Buna rağmen kopyaların uzunluğu 48 ile 79 bp arasında değişir.

4.4.2 Görsel analiz

VLTR bölgesi için tanımlama görsel olarak doğrulanmıştır. Yerleştirilmiş SSR'ler, bölgelerindeki belirsizlik nedeni ile koleksiyona dahil edilmiştir. SSR bölgesi üç dizi farklılığı içerir: İki boşluk doldurma, bir silinme ve bir eklenme. Bu düzenli olmama durumu içinde üç tam SSR bölgesi vardır. Böylece, bu dizi VLTR analizlerinde tam olmayan SSR bölgeleri içinde bulunan tam SSR'ler için test amaçlı kullanılabilir.

4.4.3 Algoritmanın performansı ve diğer yazılımlarla karşılaştırılması

Tandem Repeats Miner yazılımı 45. ve 70. indisler arasında 13 tane ard arda 2

dizilim uzunluğunda ardışık tekrar bulmaktadır. Motif GT şeklindedir. Aynı motifi 282. ve 329. indisler arasında 24 kere; 226. ve 251. indisler arasında 13 kere bulmaktadır. Bütün tekrarlar tam ardışık tekrardır. Consensus motifle birebir uyum göstermektedir. Benzer şekilde 96. ve 148. indisler arasında 27 tane ard arda 2 dizilim uzunluğunda ardışık tekrar bulmaktadır. Motif TG şeklindedir. Toplam 4 eşleşmeme durumu ve 2 adet InDel bulunmaktadır. Tandem Repeats Finder (TRF) (Benson, 1999) yazılımı aynı sonucu tespit etmiştir. TRF; TG motifini 27,5 kere bulmaktadır. (Hauth ve Joseph, 2002) aynı SSR'leri 5 farklı SSR bölgesinde GT motifi şeklinde bulduklarını belirtmiştir.

Tandem Repeats Miner yazılımı 132. ve 287. indisler arasında 3 tane ard arda 54 dizilim uzunluğunda ardışık tekrar bulmaktadır. Bir adet eşleşmeme durumu ve 3 adet Indel bulunmaktadır. TRF aynı motifi 3,3 kere bulmaktadır. (Hauth ve Joseph, 2002) SSR bulunmayan kısımlarda, uzunluğu 48 bp ile 79 bp arasında değişen VLTR bulduklarını belirtmiştir.

4.5 GenBank Lokus – DMPUGDMG1

Bu dizinin alındığı tür *Drosophila melanogaster* (meyva sineği) dir.

4.5.1 Genbank dizi bilgileri

Meyva sineği genomunda göz rengini belirleyen enzimi kodlayan dizidir. Dizi iyi korunmuş TCTCTCT motifli ardışık tekrarları içerir.

4.5.2 Görsel analiz

Dizi yüksek oranlı C/T değişimini içerir. Bir uçta CT motiflerini içeren kısa bir SSR bölgesi vardır. Sonra, 25 kopyalı tam, ardışık tekrarlı TCTCTCT motifi devam eder. Sonunda bu motif T, C ve ayrıca belirsiz herhangi bir nükleotid içeren karmaşık bir görünüm kazanır.

4.5.3 Algoritmanın performansı ve diğer yazılımlarla karşılaştırılması

Tandem Repeats Miner yazılımı 2207. ve 2416. indisler arasında 30 tane ard arda 7 dizilim uzunluğunda ardışık tekrar tespit etmektedir. Consensus motif TCTCTCT şeklindedir. Toplam 3 adet eşleşmeme durumu tespit edilmiştir. Tandem Repeats Finder (TRF) (Benson, 1999) yazılımı aynı consensus motifi belirlemiştir. TRF; aynı motifi 30,6 kere bulmaktadır. (Hauth ve Joseph, 2002) aynı motifi tam ardışık tekrar olacak şekilde 25 kere bulduğunu belirtmiştir.

4.6 GenBank Lokus – ECTRNYSU

Bu dizinin alındığı tür *Escherichia coli* (bakteri) dir.

4.6.1 Genbank dizi bilgileri

Dizi tyrT operonunu içerir. Bu geni hemen izleyen ardışık tekrarlı dizi 178 bp uzunluğunda motife sahiptir ve üç kopyası vardır.

4.6.2 Görsel analiz

Üç kopyalı 178 motif uzunluğunu hafifçe aşan ardışık tekrarlı durum bu dizi için doğrulanmıştır. Dizi içine yerleştirilmiş SSR bölgeleri ACC motifine sahiptir. Her SSR bölgesi aynı diziye sahiptir. Üç mükemmel kopya yanlardan mükemmel olmayan kopyalarla kuşatılmıştır.

4.6.3 Algoritmanın performansı ve diğer yazılımlarla karşılaştırılması

Tandem Repeats Miner yazılımı 625. ve 980. indisler arasında 2 tane ard arda 178 dizilim uzunluğunda ardışık tekrar bulmaktadır. Consensus motif 177 dizilim uzunluğundadır. 2 tane ekleme yapılmak suretiyle 178 dizilim uzunluğu elde edilmiştir. Toplam 5 adet eşleşmeme durumu tespit edilmiştir. Tandem Repeats Finder (TRF) (Benson, 1999) yazılımı aynı consensus motifi ve aynı dizilim uzunluğunu belirlemiştir. TRF; aynı motifi 2,3 kere bulmaktadır. (Hauth ve Joseph, 2002) 178 bp uzunluğunda 3 kopya bulduğunu belirtmiştir.

4.7 GenBank Lokus – HSVDJSAT

Bu dizinin alındığı tür *Homo sapiens* (insan) dir.

4.7.1 Genbank dizi bilgileri

Bu dizi yakın ilişkili 9 ve 10 bp motif uzunluğunda 36 birleşik motiften oluşmuştur. Üç ana motif **CTGGGAGAGG**, **CTGGGAGAG** ve **CTGGGATTG**'dir ve sırasıyla **1**, **2** ve **3** simgeleri ile gösterilmiştir. **1 2 1 3 1 2 1 2 1 3 1** in oluşturduğu 11 kopyalı motif birleşik motifi oluşturmuştur. Ayrıca bölge GCTGGTGG motifi ile yanlardan kuşatılmıştır.

4.7.2 Görsel analiz

Bu bölgenin ana motifleri ve motif yapıları görsel olarak ta doğrulanmıştır.

4.7.3 Algoritmanın performansı ve diğer yazılımlarla karşılaştırılması

Tandem Repeats Miner yazılımı 826. ve 856. indisler arasında 16 tane ard arda 2 dizilim uzunluğunda tam ardışık tekrar bulmaktadır. Motif AC şeklindedir ve bir adet ekleme yapılmıştır. TRF ve (Hauth ve Joseph, 2002) aynı tekrarı belirtilen sayıda bulduklarını belirtmiştir.

Tandem Repeats Miner yazılımı 1190. ve 1531. indisler arasında 18 tane ard arda 19 dizilim uzunluğunda ardışık tekrar bulmaktadır. Consensus motif yine 19 dizilim uzunluğundadır. 33 adet eşleşmeme durumu ve 21 adet indel tespit edilmiştir. TRF aynı motifi 18,4 kere bulmaktadır.

4.8 GenBank Lokus – MM102B5

Bu dizinin alındığı tür *Mus musculus* (fare) dur.

4.8.1 Genbank dizi bilgileri

Bu dizi gamma uydu bölgesinin bir parçasını içerir. Gamma uydu bölgesi 234 bp uzunluğunda ardışık tekrarlı bölgedir. 234 bp lik bölge 116 ve 118 bp uzunluğunda iki alt birimden oluşmaktadır. Bunlarda muhtemelen 9 bp uzunluğundaki 3 diziden, **GAAAAATGA, GAAAAAACT,** ve **GAAAAACGT** den oluşmaktadır. Daha ayrıntılı olarak, 234 bp uzunluğundaki gamma uydu motifi 8 altbirimden $\alpha_1\beta_1$ $\alpha_2\beta_2$ $\alpha_3\beta_3$ $\alpha_4\beta_4$ oluşmuştur. Burada α altbirimi 28 bp ye, β altbirimi 30 bp ye sahiptir.

4.8.2 Görsel analiz

58 bp motif uzunluğuna sahip ardışık tekrar bölgesi, dizi içinde tekrar eder ve yaklaşık olarak gamma uydu bölgesinin dörtte birini temsil eder. Bu durum α ve β alt birimlerinin birleşmesine uyar. Böylece bu bölge, birleşik ardışık tekrar bölgesidir.

4.8.3 Algoritmanın performansı ve diğer yazılımlarla karşılaştırılması

Tandem Repeats Miner yazılımı 1. ve 695. indisler arasında 3 tane ard arda 232 dizilim uzunluğunda ardışık tekrar bulmaktadır. Toplam 20 adet eşleşmeme durumu ve 4 adet InDel tespit edilmiştir. Tandem Repeats Finder (TRF) (Benson, 1999) yazılımı hemen hemen aynı tekrarı bulmak üzere dizilim uzunluğunu 231 olarak belirlemiştir. (Hauth ve Joseph, 2002) tekrarı 234 dizilim uzunluğunda belirlemiştir. Çok sayıda eşleşmeme durumu ve Indel tespit edilmiştir.

4.9 GenBank Lokus – MMMSAT5

Bu dizinin alındığı tür *Mus musculus* (fare) dur.

4.9.1 Genbank dizi bilgileri

Bu dizi 270 bp uzunluğunda ardışık tekrar bölgesi içerir. Bu bölge içinde bulunan SSR bölgesi **AC, AT** ve **GT** motiflerinin karışımını içerir.

4.9.2 Görsel analiz

270 bp uzunluğundaki bölge iki nükleotidli motifleri içeren pek çok Ardışık Tekrar bölgesinden oluşur. Bu motifler tek tek saptanabilir.

4.9.3 Algoritmanın performansı ve diğer yazılımlarla karşılaştırılması

Tandem Repeats Miner yazılımı 251. ve 292. indisler arasında 21 tane ard arda 2 dizilim uzunluğunda tam ardışık tekrar bulmaktadır. Motif TG şeklindedir. TRF; (Hauth ve Joseph, 2002) aynı tekrarı belirtilen sayıda bulduklarını belirtmiştir.

Tandem Repeats Miner yazılımı 118. ve 215. indisler arasında 2 tane ard arda 49 dizilim uzunluğunda ardışık tekrar bulmaktadır. Toplam 10 adet eşleşmeme durumu tespit edilmiştir. TRF; aynı tekrarı 2,2 kere bulmaktadır.

Tandem Repeats Miner yazılımı 65. ve 155. indisler arasında 5 tane ard arda 18 dizilim uzunluğunda ardışık tekrar bulmaktadır. Toplam 12 adet eşleşmeme durumu ve 2 adet Indel tespit edilmiştir. TRF; aynı tekrarı 5,4 kere bulmaktadır.

(Hauth ve Joseph, 2002) 2 dizilim uzunluğuna sahip pek çok SSR bulduğunu belirtmiştir. Tandem Repeats Miner adı geçen SSR'leri tespit edebilmektedir.

4.10 GenBank Lokus – U00144

Bu dizinin alındığı tür *Mus musculus* (fare) dur.

4.10.1 Genbank dizi bilgileri

Bu dizi çeşitli boyutlarda motifleri içeren SSR demetlerine sahiptir. Aynı motif uzunluğuna sahip birkaç SSR yer almaktadır.

4.10.2 Görsel analiz

SSR karışımları ve motif yapıları görsel olarak ta doğrulanmıştır.

4.10.3 Algoritmanın performansı ve diğer yazılımlarla karşılaştırılması

Tandem Repeats Miner yazılımı 321. ve 395. indisler arasında 37 tane ard arda 2 dizilim uzunluğunda ardışık tekrar bulmaktadır. Toplam 8 adet eşleşmeme durumu tespit edilmiştir. TRF; aynı tekrarı 37,5 kere bulmaktadır.

(Hauth ve Joseph, 2002) 2,4,6 dizilim uzunluğuna sahip pek çok SSR bulduğunu belirtmiştir. Tandem Repeats Miner adı geçen SSR'leri tespit edebilmektedir.

5. SONUÇ VE ÖNERİLER

3. Bölümde yazılım altyapısını oluşturan algoritmalara yer verildi, sonrasında daha önce geliştirilmiş algoritmalar eksi ve artı yönleri bakımından değerlendirildi, son olarak, Tandem Repeats Miner yazılımının nasıl geliştirildiği ve kullanımına ilişkin bilgiler verildi. 4. Bölümde elde edilen sonuçlar; gerek önceki çalışmalarla gerekse görsel analizle mukayese edilerek bir takım bulgular elde edildi.

Bu bölümde ise, tasarlanan ardışık tekrar arama yazılımı değerlendirilmiştir. Geliştirilen yazılımın sonuçlarına ve bu sonuçlara bağlı olarak gelecek çalışmalar için önerilere yer verilmiştir.

5.1 Sonuçlar

Bu çalışma ile, isteğe bağlı dizilim uzunluğu ayarlanabilen ardışık tekrarlı DNA dizilerini bulmaya yönelik bir yazılım geliştirilmiştir. Geliştirilen uygulamanın yazılım kodları ve çalıştırılabilir versiyonu, CD ortamında Ek-2 olarak verilmiştir.

Geliştirme esnasında, önceki çalışmalarla karşılaştırma yapılabilmesi için gen bankası dizi koleksiyonu oluşturulmuştur. Daha sonra her bir türe ait dizide consensus motif belirlenmeye çalışılmıştır. Ardışık tekrar uzunluğu saptanarak dizi baştan sona taranmıştır. Eşleşmeme durumları ve ekleme – çıkarma göz önüne alınarak ilgili motiflerin consensus motifle olan benzerliği araştırılmış ve bunların daha sonra tekrar sayısına ilave edilmesi sağlanmıştır. Veriler FASTA formatında kabul edilmiş; dizinin taranması tamamlandığında yeni dizi için aynı işlemlerin tekrarlanmasına olanak sağlanmıştır.

Elde edilen sonuçlar önceki çalışmaların sonuçlarıyla mukayese edilerek performans analizi yapılmıştır. Time complexity (zaman karmaşıklığı) üzerinde durulmuş; yazılımın çalıştırılma hızı optimize edilmeye çalışılmıştır. Bir diğer üzerinde durulan konu eşleşmeme durumları ve ilgili motiften ekleme – çıkarma yapılmasıdır. Motifler arası mesafe dikkate alınarak bu sorunun üstesinden gelinmeye çalışılmıştır. Yazılım çıktısının; önceki çalışmalarda gözlemlendiği gibi kullanıcıya tekrar sayısı, tekrarın uzunluğu, tam eşleşme durumu, eşleşmeme ve ekleme – çıkarma hususunda yeterince bilgi vermesi amaçlanmıştır.

Geliştirilen bu sistemin başarısının arttırılması ile konuyla ilgili araştırma yapan akademisyenlerin, yeni genetik mahsül oluşturma çabası içinde olan üreticilerin ve ilgili genetik hastalıkların saptanmasında çözüm üreten uzmanların gerekli materyal ve veriyi toplaması sağlanabilir. Ayrıca nedeni tam olarak belirlenemeyen genom hatalarının çözümlenmesinde araç olarak kullanılabilir.

5.2 Öneriler

Yazılım performansı, ele alınan gen bankası koleksiyonunun büyüklüğüne bağlı değişmektedir. Örneğin; binlerce karakter uzunluğundaki DNA dizilerinde motif arama hızı oldukça düşmektedir. Sistem donanımı ve işlemci hafızası da arama hızını doğrudan etkilemektedir. Bu nedenle önerilebilecek yaklaşım, yazılıma veri halinde girilen türe özgü DNA dizilerinin uzunluğunun öncelikle belli limitler dahilinde olması ve kullanılan algoritmanın gelişim sürecine bağlı olarak arttırılmasıdır.

Mevcut sorunlardan biri de eşleşmeme durumları ve ekleme – çıkarma durumlarında elde edilen kopyaların belirlenen consensus motife ne oranda benzediği ile ilgilidir. Arama algoritmasının belirlenmesinde farklı matematik ve olasılık yaklaşımları getirilerek bu sorunun üstesinden gelinebilir. Bu yaklaşımlardan biri de diğer bilim dallarında sıkça kullanılan sinyal işleme tekniğidir. Bioinformatik bilim dalına bu yeni yaklaşımların eklenmesiyle geliştirilecek yazılımın çok daha yüksek çıktılı ve hızlı olması sağlanabilir. Gerek ortak motifin belirlenmesi, gerekse eşleşmeme ve ekle – sil durumları için yeni tekniklerin uygulanması ile daha başarılı sonuçlar alınabilir.

KAYNAKLAR

Abajian, C. (1994) Sputnik. http://abajian.net/sputnik/

Ahn, S., Anderson, J.A., Sorrells, M.E. and Tanksley, S.D. (1993) Homoeologous relationships of rice, wheat and maize chromosomes. **Mol. Gen. Genet.** 241:483-490.

Baldi, P., Brunak, S., Chauvin, Y. and Pedersen, A. G. (1999) Syructural basis for triplet repeat disorders: a computational analysis. **Bioinformatics 15**(11): 918-929.

Bennetzen, J.L. and Freeling, M. (1993) Grasses as a single genetic system: Genome composition, collinearity and compatibility. **Trends Genet.** 9:259-261.

Bennetzen, J.L. and Freeling, M. (1997) The unified grass genome: synergy in synteny. **Genome Res.** 7:301-306.

Benson, G. (1999) Tandem repeats finder: a program to analyze DNA sequences. **Nucleic Acids Res.** 27:573-580.

Bilgen M., Karaca M., Onus A. N. and Ince A. G. 2004 A software program combining sequence motif searches with keywords for finding repeats containing DNA sequences. **Bioinformatics 20**, 3379–3386.

Bryant-Greenwood, P. (2002) Molecular diagnostics in obstetrics and Gynecology. **Clin Obstet Gynecol.** 45:605-621.

Cardle, L., Ramsay, L., Milbourne, D., Macaulay, M., marshall, D. and Waugh, R. (2000) Computational and experimental characterization of physically clustered simple sequence repeats in plants. **Genetics** 156:847-854.

Cordeiro, G.M., Casu, R., McIntyre, C.L., Manners, J.M. and Henry, R.J. (2001) Microsatellite markers from sugarcane (Saccharum spp.) ESTs cross transferable to erianthus and sorghum. **Plant Sci.** 160:1115-1123.

Cullis, C.A. (2002) The use of DNA polymorphisms in genetic mapping. **Genet Eng.** (N Y) 24:179-89.

Dodgson, J.B., Cheng, H.H. and Okimoto, R. (1997) DNA marker technology: A revolution in animal genetics. **Poultry Sci.** 76:1108-1114.

Eujayl, I., Sorrells, M.E., Baum, M., Wolters, P. and Powell, W. (2001) Assessment of genotypic variation among cultivated durum wheat based on EST-SSRs and genomic SSRs. **Euphytica** 119:39-43.

Fischetti, V., Landau, G., Schmidt, J. and Sellers, P. (1992) Identifying Periodic Occurrences of a Template with Applications to Protein Structure. In Apostolico, A., Crochemore, M., et al. (eds). Proceedings of the Third Annual Symposium on Combinatorial Pattern Matching, Lecture Notes in Computer Science. **Springer-Verlag**, Berlin, 644, 111-120.

Gusfield, D. (1997) Algorithms on strings, trees, and sequences. New **York: Cambridge University Press**, pp 117.

Hauth, A.M. and Joseph, D. A.(2002). Beyond Tandem Repeats: Complex Structures and Distance Regions of Similarity. Bioinformatics, 2002. July: 18. Supply1 1: S31-7.

Hearne, C.M., Ghosh, S. and Todd, J.A. (1992). Microsatellites for linkage analysis of genetic traits. **Trends Genet.** 8:288-294

Heslop- Harrison J. S. (2003) Tandemly repeated DNA sequences and centromeric chromosomal regions of Arabidopsis species. **Chromosome Res.** 241-253.

Jeffreys, A. J., Wilson V. and Thein S. J. (1985). Hypervariable 'minisatellite' regions in human DNA. **Nature** 314: 67-73.

Kantety, R.V., La Rota, M., Matthews, D.E. and Sorrells, M.E. (2002) Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum and wheat. Plant Mol. Biol. 48:501-510.

Karaca, M., Saha, S., Jenkins J. N., Zipf A., Kohel R. and Stelly, D. M. (2002). Simple sequence repeat (SSR) markers linked to the Ligon Lintless (Li1) mutant in cotton. **J. Heredity** 93: 221-224.

Keniry, M.A. (2000) Quadruplex structures in nucleic acids. **Biopolimers**, 56, 123-146

Killian, A., Chen, J., Han, F., Steffenson, B. and Kleinhofs, A. (1997) Towards map-based cloning of the barley stem rust resistance gene Rpg1 and rpg4 using rice as a intergenomic cloning vehicle. **Plant Mol. Biol.** 35:187-195.

Klintschar, M. and Wiegand, P.(2003) Polymerase slippage in relation to the uniformity of tetrameric repeat stretches. **Forensic Sci. Int.**, 135, 163-166.

Kurtz, S., Choudhuri, J.V., Ohlebusch, E., Schleiermacher, C., Stoye, J. and Giegerich, R. (2001) REPuter: the manifold applications of repeat analysis on a genomic scale. **Nucleic Acids Research** 29(22):4633-4642.

Levenshtein, V.I. (1966) Binary codes capable of correcting insertions and reversals. **Soviet Physics Dokl.** 10:707-710.

MC Murray, C.T. (1999) DNA secondary structure: a common and causative factor for expansiyon in human disease. **Proc. Natl Acad. Sci.** USA, 96, 1823-1825

McCarthy, J.J. and Hilfiker, R. (2000) The use of single-nucleotide polymorphism maps in pharmacogenomics. **Nat. Biotechnol.** 18:505-508

Moore, S.S., Sargeant, L.L., King, T.J., Mattick, J.S., Georges, M. and Hetzel, D.J. (1991) The conservation of dinucleotide microsatellites among mammalian genomes allows the use of heterologous PCR primer pairs in closely related species. **Genomics** 10:654-660.

Morgante, M. and Olivieri, A.M. (1993) PCR-amplified microsatellites as markers in plant genetics. **Plant J.** 3:175-182.

Myers, E.W. and Miller, W. (1988) Optimal alignments in linear space. **Computer Applications in the Biosciences** 4:11-17.

Nave, A., Kashi, Y. And Soller, M. (1997) Minisatellite and microsatellite length variation at a complex bovine VNTR locus. **Animal Genetics** 28(1):52-54.

Needleman, S.B. and Wunsch, C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. **Journal of Molecular Biology** 48:443-453.

Parisi, V., Fonzo, V. D. Ve Aluf- Pentini, F. (2003) STRİNG: finding tandem repeats in DNA sequences. **Bioinformatics**, 19. 1733-1738

Paterson, A.H. (1996a) DNA Marker-Assisted crop improvement. In "Genome mapping in plants". (Paterson, A.H. ed). **R.G. Landes Co.** pp. 71-79.

Paterson, A.H. (1996b) Making Genetic Maps. In "Genome mapping in plants". (Paterson, A.H. ed). **R.G. Landes Co.** pp. 23-37.

Peakall, R., Gilmore, S., Keys, W., Morgante, M. and Rafalski, A. (1998) Cross-species amplification of Soybean (Glycine max) simple sequence repeats (SSRs) within the genus and other legume genera: implications for the transferability of SSRs in plants. **Mol. Biol. Evol.** 15:1275-1287.

Pearson, C.E. and Sinden, R.R. (1998) Trinucleotide repeat DNA structures: dynamic mutations from dynamic DNA. **Current Opinion in Structural Biology** 8(3):321-30.

Pfost, D.R., Boyce-Jacino, M.T. and Grant, D.M. (2000) A SNPshot: pharmacogenetics and the future of drug therapy. **Trends Biotechnol.** 18:334-338.

Powell, W., Morgante, M., McDevitt, R., Vendramin, G. and Rafalski, J. (1995) Polymorphic simple sequence repeat regions in chloroplast genomes: applications to the population genetics of pines. **Proc. Natl. Acad. Sci. USA**, 92:7759-7763.

Powell, W., Machray, G.C. and Provan, J. (1996) Polymorphism revealed by simple sequence repeats. **Trends Plant Sci.** 1:215-222.

Rafalski, J.A. and Tingey, S.V. (1993) Genetic diagnostics in plant breeding: RAPDs, microsatellites and machines. **Trends Genet.** 9:275-280.

Rallo, P., Tenzer, I., Gessler, C., Baldoni, L., Dorado, G. and Martin, A. (2003) Transferability of olive microsatellite loci across the genus Olea. **Theor. Appl. Genet.** 107:940-946.

Reddy, P. S. and Housman, D.E. (1997) The complex pathology of trinucleotide repeats, Curr, **Opin.Cell Biol.**, 9, 364-372

Schmidt, T., Schleef, M., Friehs, K., Flaschel, E. (2003), Production of supercoiled multimeric plasmid DNA for biopharmaceutical application. J. **Biotechnol.** 105, 205-213.

Scott K. D., Eggler P., Seaton G., Rossetto M., Ablett E. M., Lee L. S. and Henry R. J. (2000) Analysis of SSRs derived from grape ESTs. **Theor. Appl. Genet.** 100, 723–726.

Shafer , R.H. and Smirnov, I. (2000) Biological aspects of DNA/RNA quadruplexes. **Biopolimers**, 56, 209-227

Sinden, R.R., Potaman, V.N., Oussatcheva, E.A., Pearson, C.E., Lyubchenko, Y.L. and Shlyakhtenko, L.S. (2002) Triplet repeat DNA structures and human genetic disease: dynamic mutations from dynamic DNA. **J. Biosci.** 27:53-65.

Smith, T. F. and Waterman, M.S. (1981) İdentification of common molecular subsequences. **Journal of Molecular Biology** 147: 195-197.

Sreenu. V. B., Vishwwanath, A., Nagaraju, J. ve Nagarajaran, H.A.(2003) MICdb: database of proprokaryotic microsatellites. **Nucleic Acids Res.**, 31, 106-108.

Tran. T. T., Emanuella II. V. A., ve Zhou G. T., " Techniques for detecting approximate tandem repeats in DNA" **. Proceedings of the İnternational Conference for Acoustics, Speech and Signal Processing (ICASSP)**, Montreal, Canada, May 2004, vol.5, pp. 449- 452.

Tautz, D. and Renz, M. (1984) Simple sequences are ubiquitous repetitive components of eukaryotic genomes. **Nucl Acids Res.** 12:4127-4138.

Terauchi, R. and Konuma, A. (1994) Microsatellite polymorphism in Dioscorea tokoro, a wild yam species. **Genome** 37:794-801.

Thiel, T., Michalek, W., Varshney, R.K. and Graner, A. (2003). Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (Hordeum vulgare L.). **Theor. Appl. Genet.** 106:411-422.

Timchenko , L.T. and Caskey, C.T. (1999) Triplet repeat disorders: discussion of molecular mechanism **Cell. Mol. Life Sci.**, 55, 1432-1447

Toth, G., Gaspari, Z. and Jurka, J. (2000) Microsatellites in different eukaryotic genomes: survey and analysis. **Genome Res.** 10:967-981.

WEB_1. (2006). Wikipedia, **,** the free encyclopedia. http://en.wikipedia.org/wiki/EST (06.05.2006).

WEB_2. (2006). http://www.genet.sickkids.on.co/~ali/repeatfinder.html (10.04.2006).

Westman, A.L. and Kresovich, S. (1998) The potential for cross-taxa simple-sequence repeat (SSR) amplification between Arabidopsis thaliana L. and crop brassicas. **Theor. Appl. Genet.** 96:272-281.

EKLER

Ek-1 TANIMLAR

AFLP: Amplified Fragment Length Polimorphism (çoğaltılmış fragmentlerin uzunluk polimorfizmi)

Alel: Kromozomun belli bir yerinde görülebilen, bir genin değişik formları

Bp: Base pair (baz çifti)

cDNA: Komplementer deoksiribonükleik asit. Tamamlayıcı DNA. Haberci RNA şablonundan sentezlenerek elde edilen DNA şeklinde de tanımlanabilir.

DAMD-PCR: PCR da miniuyduların DNA dan doğrudan çoğaltılması tekniği

dATP : deoksi adenozin trifosfat

dCTP : deoksi sitozin trifosfat

dGTP: deoksi guonozin trifosfat

dTTP: deoksi timidin trifosfat

Deoksiribonükleik asit, **DNA:** Kromozomlarda bulunur ve nükleotitlerdeki özel dizilerde kodlanan genetik bilgi içerir.

Dikotiledon: Çift çenekli bitki Embriyosunda iki çenek yaprağı bulunan bitki.

Gen Haritalaması: Bir DNA molekülündeki genlerin göreceli konumlarının belirlenmesi. Bu haritalamada hangi genin bir diğerine göre molekülün neresinde yer aldığı ve aralarında neler bulunduğu belirlenir.

Gen kodlama bölgesi: Gen kodlama bölgesi DNA nın bir parçasıdır ve mRNA ya kopyalanır ve proteine dönüştürülür.

Genom: Bir organizmanın asıl kalıtsal yapısı, gen çeşitleri

Genomics: Gen ve fonksiyonları ile ilgili çalışmalar

Genus: Yakın akraba türlerin bir araya gelerek oluşturduğu taksonomik kategori

Hairpin: DNA veye RNA nın bitişik segmentlerinin birbiri üzerine katlanması ile oluşan yapı, baz çifti ile dengede kalır.

ISSR: Basit ardışık tekrarlar arası

Kb: kilo base

Markır (Marker) : Kolaylıkla fark edilebilen DNA dizisi . Kalıtımın izlenmesi ve gen haritalarının geliştirilmesinde kullanılır.

MAS: Marker – Asisted Selection. DNA markırlarını kullanarak populasyonda arzu edilen bireylerin seçimi. Moleküler markırlar arzu edilen özelliklerle bağlantı halindedir.

Mesajcı RNA (mRNA): Nükleusta sentez edilip sitoplazmadaki ribozomlara geçen özel bir RNA çeşidi: ribozomlardaki RNA ile birleşir ve bir enzim ya da diğer bazı özel protein sentezleri için kalıp görevi yapar; elçi RNA; haberci RNA.

Nukleotid: Bir fosfat grubu, bir 5 karbonlu şeker (riboz yada dezoksiriboz) ve bir azotlu baz (pürin ya da pirimidin) dan oluşan bir molekül

Oligonücleotid: DNA veya RNA nükleotidlerinin kısa dizilimi, genellikle 20 baz çiftinden daha azdırlar.

Operon: Şifreleri tek bir mRNA molekülüne yazılan tek bir represör denetimimdeki genler

PCR : Polymerase Chain Reaction. DNA yı çoğaltma tekniği. Bu teknik ile DNA nın izole edilmesi, klonlanması ve dizi yapısı kolaylıkla gerçekleştirilir.

Polimorfizm: Biçim farklılığı

Positional cloning (Konumsal klonlama): Genleri kromozomda bulundukları konuma göre belirleyen teknik

Prion: DNA ve RNA içermeyen hastalık yapan aracı protein molekülü

Prokaryot: Zarla çevrelenmiş çekirdeği olmayan hücrelere sahip bakteri gibi tek hücreli organizmalar

RAPD: Randomly Amplified Polymorphic DNA (rastgele çoğaltılmış polimorfik DNA)

RFLP: Restriction Genetic Analysis Polymorphism (kesilmiş parçaların uzunluk polimorfizmi)

Ribonükleik asit, RNA: Riboz şekerini içeren nükleik asit. Hem nükleus hemde sitoplazmada bulunur ve protein sentezlenmesinde önemli bir moleküldür.

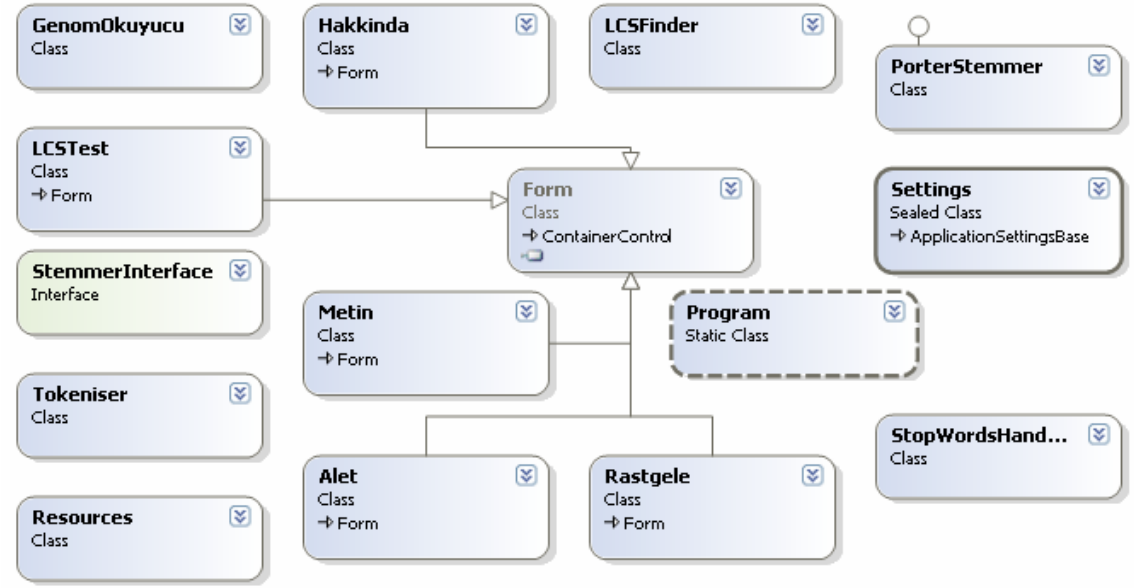
Sentromer: Her kromozomda mitosis esnasında görülen yoğunlaşmış bölge.

Telomer: Kromozomun bitiş kısmı. Bu özel yapı, doğrusal DNA moleküllerinin kendi kendini üretmesi ve dengeli yapısını koruması işlerine yarar

Transkripsiyon: Gendeki DNA dizisinin mRNA ya kopya edilmesi.

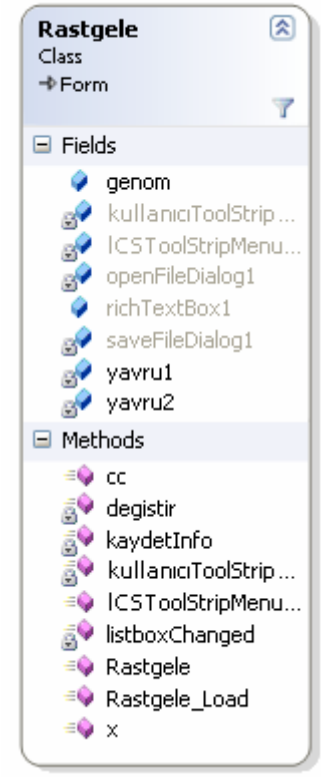
Ek-2 SINIF DİYAGRAMLARI

Aşağıda Tandem Repeats Miner yazılımına ait genel sınıf diyagramı verilmektedir. Tüm sınıf dosyaları Rastgele ana formu üzerinden çalıştırılmaktadır.



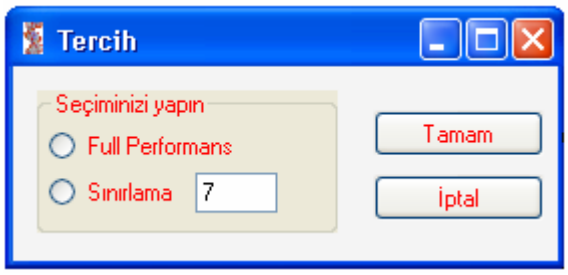
Tandem Repeats Miner genel sınıf diyagramı

Yazılımın ilk yürütülmeye başladığı sınıf; içerisinde Main( ) metodunu içeren Program statik sınıfıdır. Main( ) metodu içerisinden Rastgele ana formu çalıştırılır.

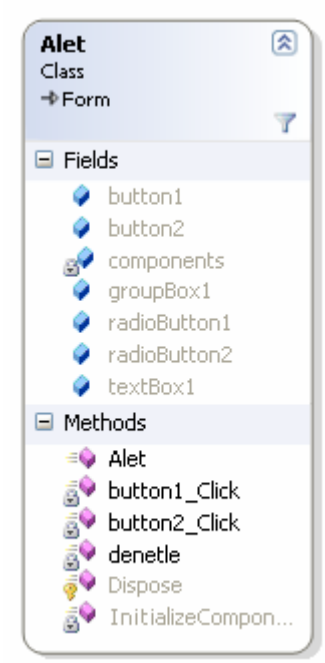


Rastgele form diyagramı

Rastgele ana formu üzerinde otomatik ardışık tekrar arama işlemi gerçekleştirilir. Otomatik arama işleminde dizi patterninin uzunluğunun belirlenme safhası Alet adı verilen form üzerinden gerçekleştirilir.



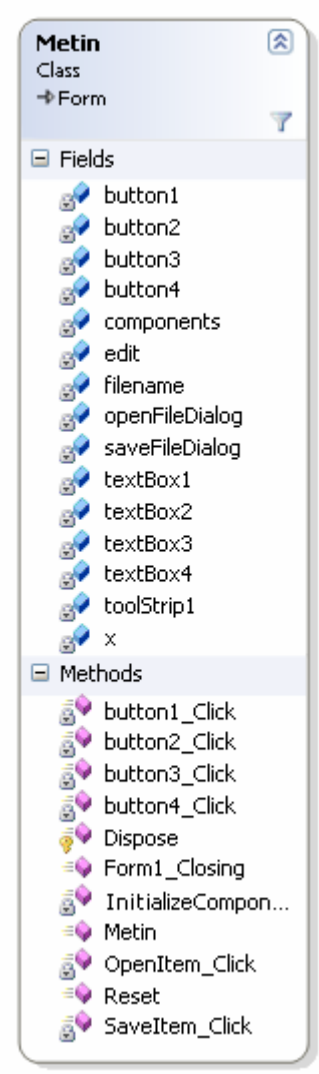
Alet formunun grafik ara yüzü



Alet form diyagramı

Yukarıdaki metodlardan button1_Click( ); belirtilen seçimi göz önüne alarak tercih verisini Rastgele ana formuna yollayarak arama işlemini başlatır. denetle( ) metodu ise; değerin girilip girilmediğini kontrol eder.

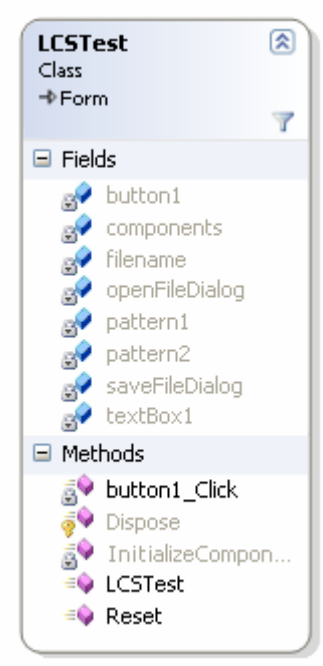
Metin adı verilen form bir yavru formdur ve Rastgele ana formu üzerinden çalıştırılır. Kullanıcı girişli olarak ardışık tekrarları ve de üst üste bindirme (overlapping) durumlarını tespit eder. Aşağıda Metin formunun sınıf diyagramı yer almaktadır.



Metin form diyagramı

Metin sınıfı içerisinde ayrıca verilerin geçerli bir format olan FASTA formatında girilmesi sağlanır. “>” işaretinden ve dizi isminden sonra ilk nükleotidden son nükleotide kadar dizi taranır; boşluklar ve FASTA’ya ait özel terimler silinir.

Bir diğer yavru form LCSTest formudur. Bu formda verilen 2 giriş dizisi arasındaki en uzun ortak alt dizi hesaplanmaktadır. Elde edilen dizi otomatik tekrar arama işleminde kullanılmaktadır. LCSTest formu 3 adet yardımcı sınıf ve 1 adet arayüzle çalışmaktadır. Yardımcı sınıflar StopWordsHandler, Tokeniser ve LCSFinder public sınıflarıdır. PorterStemmer adı verilen public sınıf StemmerInterface adlı arayüzden türetilmiştir. Aşağıda LCSTest formunun sınıf diyagramı yer almaktadır.



LCSTest form diyagramı

Belirlenen 2 veri dizisi girildikten sonra LCSTest( ) metodu kullanılarak en uzun ortak alt dizi elde edilir.

ÖZGEÇMİŞ

Onur İNAN, 1980 yılında Antalya'da doğdu. 1991 yılında Gazi Mustafa Kemal İlkokulu'nu, 1998 yılında Antalya Anadolu Lisesi'ni bitirdi. 2003 yılında Pamukkale Üniversitesi Mühendislik Fakültesi Elektrik-Elektronik Mühendisliği Bölümü'nden mezun oldu.

Eylül 2003'te Pamukkale Üniversitesi Fen Bilimleri Enstitüsü'nde yüksek lisans eğitimine başladı. Eylül 2003 – Ekim 2004 arasında iki ayrı firmada mühendis olarak görev aldı. Eylül 2004'te Süleyman Demirel Üniversitesi'nin açmış olduğu sınav neticesinde Burdur Bucak E.G.T.B.M.Y.O.'nun Endüstriyel Elektronik Programı'na Öğretim Görevlisi olarak atandı. Halen aynı kurumda görevine devam etmektedir.