



TÜRKİYE CUMHURİYETİ
KARADENİZ TEKNİK ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ

BİYOİSTATİSTİK ve TIP BİLİŞİMİ ANABİLİM DALI

**RİBOŞALTERLERİN İNSAN GENOMUNDA
ARANMASI VE BİLGİSAYAR
ORTAMINDA VAKA ÇALIŞMALARI İLE
POTANSİYEL BİR İLAÇ HEDEFİ
OLARAK TAHMİN EDİLMESİ**

Nihat Burak ZİHNİ

DOKTORA TEZİ

Prof. Dr. Kemal TURHAN

TRABZON-2023

BEYAN

Bu tez çalışmasının Karadeniz Teknik Üniversitesi Sağlık Bilimleri Enstitüsü Tez Hazırlama ve Yazım Kılavuzu standartlarına uygun olarak yapıldığını ve yazıldığını, tezin akademik ve etik kurallara bağlı kalınarak gerçekleştirilmiş özgün bir bilimsel araştırma eserim olduğunu, tezde yer alan ve bu tez çalışmasıyla elde edilmeyen tüm bilgi ve yorumlara kaynak gösterdiğimi ve kullanılan kaynakların kaynaklar listesinde yer aldığını, tezin çalışılması ve yazımı aşamalarda patent ve telif haklarını ihlal edici bir davranışımın olmadığını beyan ederim.

26.03.2023
Nihat Burak ZİHNİ
(İmza)

İthaf

Bu doktora tezimi, benim bu günlere gelmem için hiçbir fedakârlıktan kaçınmayan ve her zaman yanımda olan sevgili anneme, eşime ve biricik oğlum Celal Buğra'ya ithaf ediyorum.



İÇİNDEKİLER

Sayfa

İÇ KAPAK SAYFASI

ONAY

BEYAN

İthaf

TEŞEKKÜR

İÇİNDEKİLER

viii

TABLolar DİZİNİ

xiii

ŞEKİLLER DİZİNİ

xv

RESİMLER DİZİNİ

xvii

KISALTMALAR, SİMGELER ve FORMÜLLER DİZİNİ

xxiii

ÖZET

xxiv

ABSTRACT

xxv

1. GİRİŞ ve AMAÇ

1

2. GENEL BİLGİLER

3

2.1. Genel Bilgilerin Veriliş Şekli Hakkında Kısa Açıklama

3

2.2. Biyoinformatik Veri Tabanları ve Bağlantılı Kavramlar Hakkında Bilgiler

3

2.2.1. Biyoinformatik Nedir?

3

2.2.2. Veri Tabanı Nedir?

3

2.2.3. Biyolojik Veri Tabanı Nedir?

3

2.3. Rfam Biyoinformatik Veri Tabanı ve Bağlantılı Kavramları Hakkında Bilgiler

4

2.3.1. Rfam Nedir?

4

2.3.2. Rfam Veri Tabanındaki RNA Aileleri Nasıl Oluşturuluyor?

5

2.3.3. Cis-Düzenleyici Eleman Nedir?

5

2.3.4. Kodlamayan RNA (non-coding RNA) Nedir?

5

2.3.5. Klan (Clan) Nedir?

6

2.3.6. Aile (Family) nedir?

6

2.3.7. Kovaryans Modeli (CM) Nedir?

6

2.3.8. Infernal Yazılımı Nedir?

6

2.3.9. Tohum Hizalama (*Seed Alignment*) Nedir?

6

2.3.10. Tam Hizalama (*Full Alignment*) Nedir?

7

2.3.11. Aileye Dahil Etme Eşik Değeri (*Gathering Cutoff*) Nedir?

7

2.4. Biyoinformatik Alanına Yönelik Temel Kavramlar	7
2.4.1. RNA ve mRNA Nedir?	7
2.4.2. BLAST ve FASTA Nedir?	7
2.4.3. GFF Dosyası Nedir?	7
2.4.4. PDB ve Idx Dosyası Nedir?	8
2.4.5. İntron ve Ekzon Nedir?	8
2.4.6. <i>Gap</i> , <i>Gap Penalty</i> ve E-değeri Nedir?	8
2.4.7. GC Oranı Nedir?	9
2.4.8. <i>Annotation</i> Nedir?	9
2.4.9. <i>Pairwise Alignment</i> Nedir?	9
2.4.10. <i>Secondary Structure Prediction</i> Nedir?	9
2.4.11. <i>Sunburst</i> Grafiği Nedir?	9
2.5. Bilgisayar ve Kodlama Terminolojisiyle İlgili Kavramlar	9
2.5.1. <i>Phyton</i> ve <i>PyCharm</i> Nedir?	9
2.5.2. <i>Anaconda</i> ve <i>PyPI</i> Nedir?	10
2.5.3. SQL ve MYSQL Nedir?	10
2.5.4. Web Servisi ve Api Nedir?	11
2.5.5. Pipe-Line (Ardışık Düzen) Nedir?	11
2.5.6. <i>Regular Expression</i> (Düzenli İfade) Nedir?	11
2.5.7. TRUBA ve Slurm Nedir?	11
2.5.8. Çalışmada Kullanılan Çeşitli Dosya Türleri Hakkında Açıklamalar	12
2.6. Sanal Tarama Analiziyle İlgili Kavramlar	12
2.6.1. Ligand Nedir?	12
2.6.2. Reseptör Nedir?	12
2.6.3. <i>Docking</i> (Moleküler Yanaştırma) Nedir?	13
2.6.4. <i>Blind Docking</i> Nedir?	13
2.6.5. AD4 Nedir?	13
2.6.6. <i>AutoDock Vina</i> Nedir?	13
2.6.7. BFGS Algoritması Nedir?	13
2.6.8. <i>Exhaustiveness</i> Nedir ve Neden Önemlidir?	14
2.6.9. <i>PyMOL</i> , <i>PDBQT</i> ve <i>MOL2</i> Nedir?	14
2.6.10. <i>Virtual Screening</i> (VS) Nedir?	14
2.6.11. <i>Lipinski</i> 'nin Beş Kuralı Nedir?	14

2.6.12. Tanimoto Benzerliği Nedir?	15
2.7. İstatistik Analizleriyle İlgili Kavramlar	15
2.7.1. Tek Örneklem T-Testi Nedir?	15
2.7.2. NLP (Doğal Dil İşleme) Nedir?	15
2.7.3. LDA (Doğrusal Ayırma Analizi) Nedir?	15
2.7.4. <i>Wordcloud</i> (Kelime Bulutu) Nedir?	16
2.7.5. <i>Text Mining</i> (Metin Madenciliği) Nedir?	16
2.7.6. <i>ICC</i> (Sınıf İçi Korelasyon Katsayısı) Nedir?	16
3. GEREÇ VE YÖNTEM	17
3.1. Veri Seti Derleme Prosedürü Hakkında Açıklama	17
3.2. Geliştirilen Ardışık Düzen Hakkında Ön Bilgilendirme	18
3.3. <i>Rfam</i> Veri Tabanı Kısmı İçin Oluşturulan Kod Betikleri	19
3.3.1. <i>Rfam</i> Veri Setinin Oluşturulması	20
3.3.2. Riboşalter RNA Ailesine Ait Üyelerin <i>Rfam</i> Biyolojik Veri Tabanına Özel Kimlik Numaralarının Tespit Edilmesi	22
3.3.3. Riboşalter RNA Ailesine Ait Üyeler İçin Hizalama Veri Setinin Oluşturulması	23
3.3.4. Riboşalter RNA Ailesine Ait Üyeler İçin FASTA Dosyalarının İndirilmesi	25
3.3.5. Riboşalter RNA Ailesine Ait Üyeler İçin GC-Oranı Analizlerinin Yapılması	27
3.3.6. Riboşalter RNA Ailesine Ait Üyelerin Hizalanmış Sekanslarının İnsan Genomunda Aranması	31
3.3.6.1. NCBI Biyolojik Veri Tabanı Üzerinden BLAST Analizi Süreci	31
3.3.6.2. RNA Central Biyolojik Veri Tabanı Üzerinden “ <i>Anotasyon</i> ” Analizi Süreci	35
3.3.7. İnsan (<i>H. sapiens</i>) Genomuna Ait GFF Dosyası Üzerinden Anotasyon Bölgelerinin Analizi	42
3.3.8. <i>Rfam</i> Biyolojik Veri Tabanında Bulunan Taksonomik Veri Setinin Derlenmesi ve “ <i>Sunburst</i> ” Grafik Çıktılarının Elde Edilmesi	53
3.3.9. <i>Rfam</i> Biyolojik Veri Tabanında Bulunan Literatür Referansları Veri Seti Üzerinden Tanımlayıcı Grafik Çıktılarının Elde Edilmesi	61
3.3.10. <i>Rfam</i> Biyolojik Veri Tabanında Bulunan Literatür Referansları Veri Seti Üzerine Metin Madenciliği Uygulanarak Grafik Çıktılarının Elde Edilmesi	70
3.3.11. <i>Rfam</i> Biyolojik Veri Tabanında Bulunan Üç Boyutlu Yapı Veri Seti İçeriğinde Yer Alan Pdb Dosyalarının İndirilmesi	75

3.3.12. Rcsb Biyolojik Veri Tabanından Elde Edilen İlave Veri Dosyalarıyla Veri Zenginleştirilmesi Yapılması	79
3.3.13. Reseptör Adayı Üç Boyutlu Yapı Kayıtlarını Tutan Dosyanın Sonraki Süreçler İçin Raporlanması	82
3.4. ZINC Veri Tabanı Kısmı	87
3.4.1. ZINC Biyolojik Veri Tabanı Statik Dosyalarının Yerel Diske İndirilmesi	90
3.4.2. İlaç Adayı Molekül Havuzu Dosyalarının Elde Edilmesi Amacıyla Gerekli Ön Veri İşlemesi	92
3.4.3. Sanal Tarama Analizleri İçin Gerekli İlaç Adayı Molekül Dosya Havuzunun Oluşturulması	94
3.4.4. İlaç Adayı Molekül Dosya Havuzunu Oluşturan ZINC Katalog Verilerine Ait Fiziksel Özellik Uzayı Veri Sıklıklarının Hesaplanması ve Grafiklerinin Çizdirilmesi	97
3.5. Sanal Tarama Analizi Kısmı	100
3.5.1. Sanal Tarama Analizi İçin Aday Reseptörler Arasından Dosya Seçimi	101
3.5.2. Sanal Tarama Analizi İçin Gerekli Girdi Dosyalarının Oluşturulması	105
3.5.3. Sanal Tarama Analizi Öncesinde “ <i>Docking</i> ” Başarısının ve Doğruluğunun Tespiti İçin Ön Analiz Yapılması	106
3.5.4. Sanal Tarama Analizlerinin Yapılması ve Sonuç Dosyalarının Elde Edilmesi	110
3.5.5. Sanal Tarama Analizi Sonuç Dosyalarının Raporlanması	112
4. BULGULAR	113
4.1. Riboşalter Aile Üyesi RNA Elemanların Taksonomik Dağılımları	113
4.2. GC-Oranı Analizi İçin Elde Edilen İstatistiksel Sonuç Dağılımları	114
4.3. Reseptör Adayı Riboşaltere Ait Üç Boyutlu Yapı Dosyalarının Taksonomik Dağılımı	115
4.4. RNA Central Biyolojik Veri Tabanı Üzerinden “Anotasyon” Analizi Süreci Sonunda Elde Edilen Bulgular	115
4.5. ZINC Veri Tabanından Çekilen İlaç Adayı Molekül Dosyalarına Ait Çeşitli Özellik Dağılımları	116
4.6. ZINC Veri Tabanından Çekilen İlaç Adayı Molekül Dosyalarının “ <i>Lipinski’nin Beş Kuralına</i> ” Göre Dağılımı	118
4.7. <i>V. vulnificus</i> Kaynaklı Hastalık Tedavisinde Kullanılan İlaçlar ve Diğer Moleküllerle Olan <i>Tanimoto</i> Benzerlikleri	119
4.8. Sanal Tarama Analizlerinin Doğruluğunun Belirlenmesi	119

4.9. <i>Lipinski'nin Beş Kuralı</i> Dikkate Alındığında En İyi Bağlanma Enerjisine Sahip İlk On Molekül	121
4.10. <i>Lipinski'nin Beş Kuralı</i> Dikkate Alınmadığında En İyi Bağlanma Enerjisine Sahip İlk On Molekül	122
5. TARTIŞMA ve SONUÇ	123
6. KAYNAKLAR	126
EKLER	140
EK 1. Tez Çalışması Kapsamında Elde Edilen Dijital Veriler	141
	142



TABLÖLAR DİZİNİ

Tablo No	Sayfa
Tablo 1. Tez çalışmasında kullanılan çeşitli biyolojik veri tabanları	4
Tablo 2. <i>Rfam</i> üyesi her bir RNA elemanı için düzenlenen web sayfasına ait içerik türleri	4
Tablo 3. Çalışmada kullanılan dosya türleri ve açıklamaları	12
Tablo 4. Tez çalışması için geliştirilen ardışık düzen süreçlerine ait kod betikleri	18
Tablo 5. Kod betik dizinleri için kullanılan özel dosyalama tekniği içerikleri	19
Tablo 6. <i>Rfam</i> veri tabanı kısmı için oluşturulan kod betikleri	20
Tablo 7. ZINC veri tabanında arşivleme için kullanılan fiziksel özellik uzayı	89
Tablo 8. <i>Rfam</i> veri tabanı kısmı için oluşturulan kod betikleri	89
Tablo 9. Sanal tarama kısmı için oluşturulan kod betikleri ve kısa açıklamaları	101
Tablo 10. Reseptör aday dosyalar içerisinde tekil dosyaya sahip olan kayıtlar	104
Tablo 11. Riboşalterlerin kaynak genomlara dağılımı	113
Tablo 12. Tek örneklem t-testi p istatistiğine göre alınan karar dağılımları	114
Tablo 13. Reseptör aday riboşalterlere ait üç boyutlu yapı dosyalarının taksonomik dağılımı	115
Tablo 14. Süreç sonunda elde edilen anlamlı kabul edilebilecek sınırdaki bulgular	115
Tablo 15. Tablo 14'te verilen BLAST sorgu numaralarının açıklaması	116
Tablo 16. İlaç adayı molekül dosyalarının moleküler ağırlığına göre dağılımı	116
Tablo 17. İlaç adayı molekül dosyalarının <i>logP</i> (<i>xlogP</i>) değerine göre dağılımı	117
Tablo 18. İlaç adayı molekül dosyalarının <i>reactivity</i> değişkenine göre dağılımı	117
Tablo 19. İlaç adayı molekül dosyalarının <i>purchasability</i> değişkenine göre dağılımı	117
Tablo 20. İlaç adayı molekül dosyalarının <i>pH seviyesi</i> değişkenine göre dağılımı	118
Tablo 21. İlaç adayı molekül dosyalarının <i>net moleküler yük</i> değişkenine göre dağılımı	118
Tablo 22. Çalışmada kullanılan molekül dosyalarının RO5 uyumuna göre dağılımı	118

Tablo 23. <i>V. vulnificus</i> tedavisinde kullanılan ilaçlar ve bu ilaçlara benzer moleküllerin <i>Tanimoto</i> Benzerliği Tanımlayıcı İstatistikleri (n=1000)	119
Tablo 24. Sanal tarama analiz sonuçlarına dair sınıf içi korelasyon istatistikleri	120
Tablo 25. Sanal tarama analizi sonucunda elde edilen en iyi bağlanma enerjisi'ne sahip ilk on molekül (<i>Lipinski'nin beş kuralı dikkate alınmış</i>)	121
Tablo 26. En iyi bağlanma enerjisi'ne sahip ilk on molekül ile <i>V. vulnificus</i> hastalığı tedavisinde kullanılan ilaç molekülleri arasındaki hesaplanmış <i>Tanimoto</i> benzerlikleri (<i>Lipinski'nin beş kuralı dikkate alınmış</i>)	121
Tablo 27. Sanal tarama analizi sonucunda elde edilen en iyi bağlanma enerjisi'ne sahip ilk on molekül (<i>Lipinski'nin beş kuralı dikkate alınmamış</i>)	122
Tablo 28. En iyi bağlanma enerjisi'ne sahip ilk on molekül ile <i>V. vulnificus</i> hastalığı tedavisinde kullanılan ilaç molekülleri arasındaki hesaplanmış <i>Tanimoto</i> benzerlikleri (<i>Lipinski'nin beş kuralı dikkate alınmamış</i>)	122

ŞEKİLLER DİZİNİ

Şekil No	Sayfa
Şekil 1. DNA'daki aşağı akış ve yukarı akış bölgeleri	5
Şekil 2. <i>Rfam</i> veri tabanı oluşturma süreci için tasarlanan dizin yapısı ve elde edilen dosyalar	21
Şekil 3. Riboşalter RNA ailesine ait üyelerin <i>Rfam</i> biyolojik veri tabanına özel kimlik numaralarının tespit edilmesi süreci için tasarlanan dizin yapısı ve elde edilen dosyalar	23
Şekil 4. Riboşalter RNA ailesine ait üyeler için hizalama veri setinin oluşturulması süreci için tasarlanan dizin yapısı ve elde edilen dosyalar	25
Şekil 5. Riboşalter RNA ailesine ait üyeler için FASTA dosyalarının indirilmesi süreci için tasarlanan dizin yapısı ve elde edilen dosyalar	27
Şekil 6. Riboşalter RNA ailesine ait üyeler için GC-oranı analizi süreci için tasarlanan dizin yapısı ve elde edilen dosyalar	31
Şekil 7. Riboşalter RNA ailesine ait üyeler için BLAST analizi süreci için tasarlanan dizin yapısı ve elde edilen dosyalar	35
Şekil 8. Riboşalter RNA ailesine ait üyeler için “ <i>anotasyon</i> ” analizi süreci için tasarlanan dizin yapısı ve elde edilen dosyalar	42
Şekil 9. İnsan (<i>H. sapiens</i>) genomuna ait GFF dosyası üzerinden “ <i>anotasyon</i> ” bölgelerin analizi süreci için tasarlanan dizin yapısı ve elde edilen dosyalar	53
Şekil 10. <i>Rfam</i> biyolojik veri tabanında bulunan taksonomik veri setinin derlenmesi ve “ <i>sunburst</i> ” grafik çıktılarının elde edilmesi süreci için tasarlanan dizin yapısı ve elde edilen dosyalar	61
Şekil 11. <i>Rfam</i> biyolojik veri tabanında bulunan literatür referansları veri seti üzerinden tanımlayıcı grafik çıktılarının elde edilmesi süreci için tasarlanan dizin yapısı ve elde edilen dosyalar	70
Şekil 12. <i>Rfam</i> biyolojik veri tabanında bulunan literatür referansları veri seti üzerine metin madenciliği uygulanarak grafik çıktılarının elde edilmesi süreci için tasarlanan dizin yapısı ve elde edilen dosyalar	75

Şekil 13. <i>Rfam</i> biyolojik veri tabanında bulunan üç boyutlu yapı veri seti içeriğinde yer alan <i>pdb</i> dosyalarının indirilmesi süreci için tasarlanan dizin yapısı ve elde edilen dosyalar	79
Şekil 14. <i>Rcsb</i> biyolojik veri tabanından elde edilen ilave veri dosyalarıyla veri zenginleştirmesi yapılması süreci için tasarlanan dizin yapısı ve elde edilen dosyalar	82
Şekil 15. Reseptör adayı üç boyutlu yapı kayıtlarını tutan dosyanın sonraki süreçler için raporlanması süreci için tasarlanan dizin yapısı ve elde edilen dosyalar	87
Şekil 16. https://wiki.docking.org/index.php/Resources adresinde yer alan kaynak tablosu	88
Şekil 17. <i>ZINC</i> biyolojik veri tabanı statik dosyalarının yerel diske indirilmesi süreci için tasarlanan dizin yapısı ve elde edilen dosyalar	91
Şekil 18. İlaç adayı molekül havuzu dosyalarının elde edilmesi amacıyla gerekli ön veri işlemesi süreci için tasarlanan dizin yapısı ve elde edilen dosyalar	94
Şekil 19. Sanal tarama analizleri için gerekli ilaç adayı molekül dosya havuzunun oluşturulması süreci için tasarlanan dizin yapısı ve elde edilen dosyalar	96
Şekil 20. “İlaç Adayı Molekül Dosya Havuzunu Oluşturan <i>ZINC</i> Katalog Verilerine Ait Fiziksel Özellik Uzayı Veri Sıklıklarının Hesaplanması ve Grafiklerinin Çizdirilmesi” süreci için tasarlanan dizin yapısı ve elde edilen dosyalar	100
Şekil 21. Sanal tarama analizi için aday reseptör dosyalarından seçim yapılması ve seçilen dosya içeriğinin, analiz için hazır hale getirilmesi süreci için tasarlanan dizin yapısı ve elde edilen dosyalar	105
Şekil 22. Sanal tarama analizi için aday reseptör dosyalarından seçim yapılması ve seçilen dosya içeriğinin, analiz için hazır hale getirilmesi süreci için tasarlanan dizin yapısı ve elde edilen dosyalar	110
Şekil 23. Dosya sayı kısıtlaması nedeniyle geliştirilen yeni yöntemle TRUBA sunucu kaynağı kullanılarak sanal tarama analizinin yapılma süreci	111

RESİMLER DİZİNİ

Resim No	Sayfa
Resim 1. İkincil yapı sarmal bölgeleriyle renklendirilmiş örnek bir tohum hizalaması	6
Resim 2. <i>Rfam</i> veri tabanı oluşturma sürecinde kullanılan kod betiğinin yaptığı işlemler	21
Resim 3. Riboşalter RNA ailesine ait üyelerin <i>Rfam</i> biyolojik veri tabanına özel kimlik numaralarının tespit edilmesi sürecinde kullanılan kod betiğinin yaptığı işlemler	22
Resim 4. Riboşalter RNA ailesine ait üyeler için hizalama veri setinin oluşturulması sürecinde kullanılan kod betiğinin yaptığı işlemler	24
Resim 5. Riboşalter RNA ailesine ait üyeler için FASTA dosyalarının indirilmesi sürecinde kullanılan kod betiğinin yaptığı işlemler	26
Resim 6. GC-Oranı analizlerine ait ifade edilen işlem akışındaki ikinci adımda ifade edilen sekans uzunluğu ve sekansın okunma şekli ifadelerinin betimlenmesi	28
Resim 7. GC-Oranı analizlerine ait ifade edilen işlem akışındaki üçüncü adımda ifade edilen tek örneklem t istatistiği hesaplanma sürecinin betimlenmesi	29
Resim 8. GC-Oranı analizlerine ait ifade edilen işlem akışındaki dördüncü adımda ifade edilen “ <i>html</i> ” rapor çıktısından bir kesit	29
Resim 9. Riboşalter RNA ailesine ait üyeler için GC-oranı analizi sürecinde kullanılan kod betiğinin yaptığı işlemler	30
Resim 10. BLAST analizi sürecinde girdi olarak kullanılan “ <i>alignment_fragments.fasta</i> ” dosyasının oluşturulma süreci	32
Resim 11. BLAST analizlerine ait ifade edilen işlem akışındaki dördüncü adımda ifade edilen “ <i>html</i> ” rapor çıktısından bir kesit	33
Resim 12. Riboşalter RNA ailesine ait üyeler için BLAST analizi sürecinde kullanılan kod betiğinin yaptığı işlemler	34
Resim 13. Riboşalter RNA ailesine ait üyeler için anotasyon analizi sürecinde kullanılan kod betiğine ait akış sürecinin birinci adımında yapılan işlemler	36

Resim 14.	Riboşalter RNA ailesine ait üyeler için “ <i>anotasyon</i> ” analizi sürecinde kullanılan kod betiğine ait akış sürecinin ikinci adımında yapılan işlemler	37
Resim 15.	Riboşalter RNA ailesine ait üyeler için “ <i>anotasyon</i> ” analizi sürecinde kullanılan kod betiğine ait akış sürecinin dördüncü adımında yapılan işlemler	39
Resim 16.	Riboşalter RNA ailesine ait üyeler için “ <i>anotasyon</i> ” analizi sürecinde kullanılan kod betiğine ait akış sürecinin dördüncü adımında yapılan işlemler	40
Resim 17.	Riboşalter RNA ailesine ait üyeler için anotasyon analizi sürecinde kullanılan kod betiğine ait akış sürecinin beşinci adımında yapılan işlemler	41
Resim 18.	İnsan (<i>H. sapiens</i>) genomuna ait <i>GFF</i> dosyası üzerinden anotasyon bölgelerin analizi sürecinde kullanılan kod betiğine ait akış sürecinin birinci adımında yapılan işlemler	44
Resim 19.	İnsan (<i>H. sapiens</i>) genomuna ait <i>GFF</i> dosyası üzerinden anotasyon bölgelerin analizi sürecinde kullanılan kod betiğine ait akış sürecinin ikinci adımında yapılan işlemler	45
Resim 20.	İnsan (<i>H. sapiens</i>) genomuna ait <i>GFF</i> dosyası üzerinden anotasyon bölgelerin analizi sürecinde kullanılan kod betiğine ait akış sürecinin üçüncü adımında yapılan işlemler	47
Resim 21.	İnsan (<i>H. sapiens</i>) genomuna ait <i>GFF</i> dosyası üzerinden anotasyon bölgelerinin analizi sürecinde kullanılan kod betiğine ait akış sürecinin dördüncü adımında yapılan işlemler	49
Resim 22.	İnsan (<i>H. sapiens</i>) genomuna ait <i>GFF</i> dosyası üzerinden elde edilen anotasyon türleri frekans dağılım grafiklerin toplu olarak betimlenmesi birinci kısım	50
Resim 23.	İnsan (<i>H. sapiens</i>) genomuna ait <i>GFF</i> dosyası üzerinden elde edilen anotasyon türleri frekans dağılım grafiklerin toplu olarak betimlenmesi ikinci kısım	51
Resim 24.	İnsan (<i>H. sapiens</i>) genomuna ait <i>GFF</i> dosyası üzerinden anotasyon bölgelerin analizi sürecinde kullanılan kod betiğine ait akış sürecinin altıncı adımında yapılan işlemler	52

Resim 25.	<i>Rfam</i> biyolojik veri tabanında bulunan taksonomik veri setinin derlenmesi ve “ <i>sunburst</i> ” grafik çıktılarının elde edilmesi sürecinde kullanılan kod betiğine ait akış sürecinin birinci adımında yapılan işlemler	54
Resim 26.	<i>Rfam</i> biyolojik veri tabanında bulunan taksonomik veri setinin derlenmesi ve “ <i>sunburst</i> ” grafik çıktılarının elde edilmesi sürecinde kullanılan kod betiğine ait akış sürecinin ikinci adımında yapılan işlemler	56
Resim 27.	<i>Rfam</i> biyolojik veri tabanında bulunan taksonomik veri setinin derlenmesi ve “ <i>sunburst</i> ” grafik çıktılarının elde edilmesi sürecinde kullanılan kod betiğine ait akış sürecinin üçüncü adımında yapılan işlemler	57
Resim 28.	“ <i>all_members_faceted_bar.csv</i> ” dosyası içerisinde yer alan her bir Riboşalter aile üyesi RNA elemanının, “ <i>annotate</i> ” edildiği kaynak genomların, “ <i>superkingdom</i> ” taksonomi düzeyindeki frekans dağılımının görselleştirilmesi sürecinde elde edilen örnek bir grafik çıktısı	57
Resim 29.	Riboşalter aile üyesi RNA elemanının, “ <i>annotate</i> ” edildiği kaynak genomların, “ <i>superkingdom</i> ” taksonomi düzeyindeki frekans dağılım grafiklerinin toplu olarak betimlenmesi birinci kısım	58
Resim 30.	Riboşalter aile üyesi RNA elemanının, “ <i>annotate</i> ” edildiği kaynak genomların, “ <i>superkingdom</i> ” taksonomi düzeyindeki frekans dağılım grafiklerinin toplu olarak betimlenmesi ikinci kısım	59
Resim 31.	<i>Rfam</i> biyolojik veri tabanında bulunan taksonomik veri setinin derlenmesi ve “ <i>sunburst</i> ” grafik çıktılarının elde edilmesi sürecinde kullanılan kod betiğine ait akış sürecinin dördüncü adımında yapılan işlemler	60
Resim 32.	<i>Rfam</i> biyolojik veri tabanında bulunan literatür referansları veri seti üzerinden tanımlayıcı grafik çıktılarının elde edilmesi sürecinde kullanılan kod betiğine ait akış sürecinin birinci adımında yapılan işlemler	62
Resim 33.	<i>Rfam</i> biyolojik veri tabanında bulunan literatür referansları veri seti üzerinden tanımlayıcı grafik çıktılarının elde edilmesi sürecinde	63

kullanılan kod betiğine ait akış sürecinin ikinci adımında yapılan işlemler	
Resim 34. <i>Rfam</i> biyolojik veri tabanında bulunan literatür referansları veri seti üzerinden tanımlayıcı grafik çıktılarının elde edilmesi sürecinde kullanılan kod betiğine ait akış sürecinin üçüncü adımında yapılan işlemler	65
Resim 35. <i>Rfam</i> biyolojik veri tabanında bulunan literatür referansları veri seti üzerinden tanımlayıcı grafik çıktılarının elde edilmesi sürecinde kullanılan kod betiğine ait akış sürecinin dördüncü adımında yapılan işlemler sonucunda elde edilen grafik çıktılar	66
Resim 36. <i>Rfam</i> biyolojik veri tabanında bulunan literatür referansları veri seti üzerinden tanımlayıcı grafik çıktılarının elde edilmesi sürecinde kullanılan kod betiğine ait akış sürecinin dördüncü adımında yapılan işlemler sonucunda elde edilen grafik çıktılar	67
Resim 37. Makalelerin basım yılı verisinin RNA elemanlarına ait frekans dağılımlarını içeren grafik çıktılarının toplu gösteriminin birinci kısmından alınan kesit	68
Resim 38. Makalelerin basım yılı verisinin RNA elemanlarına ait frekans dağılımlarını içeren grafik çıktılarının toplu gösteriminin ikinci kısmından alınan kesit	68
Resim 39. Makalelerin yayınlandığı dergi isimleri verisinin RNA elemanlarına ait frekans dağılımlarını içeren grafik çıktılarının toplu gösteriminin birinci kısmından alınan kesit	69
Resim 40. Makalelerin yayınlandığı dergi isimleri verisinin RNA elemanlarına ait frekans dağılımlarını içeren grafik çıktılarının toplu gösteriminin ikinci kısmından alınan kesit	69
Resim 41. <i>Rfam</i> biyolojik veri tabanında bulunan literatür referansları veri seti üzerine metin madenciliği uygulanarak grafik çıktılarının elde edilmesi sürecinde kullanılan kod betiğine ait akış sürecinin ikinci adımında yapılan işlemler	71
Resim 42. <i>Rfam</i> biyolojik veri tabanında bulunan literatür referansları veri seti üzerine metin madenciliği uygulanarak grafik çıktılarının elde	73

	edilmesi sürecinde kullanılan kod betiğine ait akış sürecinin ikinci adımında yapılan işlemler	
Resim 43.	<i>Rfam</i> biyolojik veri tabanında bulunan literatür referansları veri seti üzerine metin madenciliği uygulanarak grafik çıktılarının elde edilmesi sürecinde kullanılan kod betiğine ait akış sürecinin üçüncü adımında yapılan işlemler	74
Resim 44.	“ <i>Rfam</i> ” biyolojik veri tabanında bulunan üç boyutlu yapı veri seti içeriğinde yer alan pdb dosyalarının indirilmesi sürecinde, kullanılan kod betiğine ait akış sürecinin birinci adımında yapılan işlemler	76
Resim 45.	“ <i>Rfam</i> ” biyolojik veri tabanında bulunan üç boyutlu yapı veri seti içeriğinde yer alan pdb dosyalarının indirilmesi sürecinde, kullanılan kod betiğine ait akış sürecinin ikinci adımında yapılan işlemler	77
Resim 46.	<i>Rcsb</i> biyolojik veri tabanından elde edilen ilave veri dosyalarıyla veri zenginleştirilmesi yapılması sürecinde, kullanılan kod betiğine ait akış süreci adımlarında yapılan işlemler	81
Resim 47.	Reseptör Adayı Üç Boyutlu Yapı Kayıtlarını Tutan Dosyanın Sonraki Süreçler İçin Raporlanması sürecinde, kullanılan kod betiğine ait akış sürecinin birinci adımında yapılan işlemler	83
Resim 48.	Reseptör Adayı Üç Boyutlu Yapı Kayıtlarını Tutan Dosyanın Sonraki Süreçler İçin Raporlanması sürecinde, kullanılan kod betiğine ait akış sürecinin ikinci adımında yapılan işlemler	84
Resim 49.	Reseptör Adayı Üç Boyutlu Yapı Kayıtlarını Tutan Dosyanın Sonraki Süreçler İçin Raporlanması sürecinde, kullanılan kod betiğine ait akış sürecinin üçüncü adımında yapılan işlemler	85
Resim 50.	Reseptör adayı üç boyutlu yapı kayıtlarını tutan dosyanın sonraki süreçler için raporlanması sürecinde, kullanılan kod betiğine ait akış sürecinin dördüncü adımında yapılan işlemler	86
Resim 51.	İlaç adayı moleküllerin indirilebilmesi için gerekli metin girdi dosya içeriğinin oluşturulma süreci	90
Resim 52.	İlaç adayı moleküllerin indirilebilmesi için gerekli “ <i>download_list</i> ” adlı dosya içeriğinin oluşturulmuş hali	91

Resim 53.	Yerel diske indirilen ham ilaç adayı molekölü dosyalarının oluşturulabilmesi için gerekli olan “ <i>raw_promoters</i> ” dizininin oluşturulma süreci	93
Resim 54.	Sıkıştırılmış arşiv dosyası “ <i>mol.gz</i> ” türündeki üç boyutlu “ <i>ready to dock</i> ” düzenindeki dosyaların yerel diske indirilmesi süreci	95
Resim 55.	Yerel diske indirilmiş olan “ <i>mol.gz</i> ” formatındaki üç boyutlu “ <i>ready to dock</i> ” düzenindeki dosyaların ligand “ <i>pdbqt</i> ” formatına dönüşüm sürecini temsil eden bir kesit	96
Resim 56.	Kod betiği ile elde edilen ZINC kataloglarına ait fiziksek özellik uzayı sıklık dağılım verileri üzerinden, tüm kataloglara ait veriyi içeren “ <i>descriptives.csv</i> ” dosyasının oluşturulması sürecinde yapılan işlem mantığını betimleyen den bir kesit	97
Resim 57.	Kod betiği ile elde edilen ZINC kataloglarına ait fiziksek özellik uzayı sıklık dağılım verileri üzerinden, sıklık dağılım grafiklerinin elde edilmesi sürecindeki işlem adımlarını temsil eden bir kesit	98
Resim 58.	ZINC veri tabanını oluşturan her bir kataloğa ait fiziksel özellik uzayı verilerini temsil eden sıklık dağılım grafiği çıktılarından bir kesit	99
Resim 59.	Sanal tarama analizi için seçilen aday reseptör dosyasında temizleme işleme yapılma süreci	103
Resim 60.	Sanal tarama analizi için aday reseptör dosyalarından seçim yapılması ve seçilen dosya içeriğinin analiz için hazır hale getirilmesi süreci	104
Resim 61.	Sanal tarama analizi için seçilen aday reseptörün kutu boyutu girdi değerlerinin hesaplanma süreci	106
Resim 62.	Sanal tarama analizi için aday reseptör dosyalarından seçim yapılması ve seçilen dosya içeriğinin analiz için hazır hale getirilmesi süreci	108
Resim 63.	“ <i>icc_data_points.csv</i> ” dosya içeriğinin oluşturulma süreci	108
Resim 64.	“ <i>icc_data_centroids.csv</i> ” dosya içeriğinin oluşturulma süreci	109
Resim 65.	Sanal tarama analizi sonuçları html rapor dosyası içeriğinden bir kesit	112
Resim 66.	Sınıf içi korelasyon analizi verilerinden örnek gösterimler	120

KISALTMALAR, SİMGELER ve FORMÜLLER DİZİNİ

Kısaltmalar

ABD	Anabilim Dalı
AD4	<i>AutoDock Tools 4th Edition</i>
BLAST	<i>Basic Local Alignment Search Tool</i>
C	Sitozin
DNA	Deoksiribo Nükleik Asit
RNA	Ribonükleik Asit
FASTA	<i>FAST-All</i>
FDA	<i>Food and Drug Administration</i>
FTP	<i>File Transfer Protocol</i>
G	Guanin
GA1	<i>Glutaric Aciduria Type 1</i>
GFF	<i>General Feature Format</i>
ICC	<i>Intraclass Correlation Coefficient</i>
KTÜ	Karadeniz Teknik Üniversitesi
LDA	<i>Linear Discriminant Analysis</i>
MDR	<i>Multidrug-Resistant</i>
mRNA	Mesajcı Ribonükleik Asit
NCBI	<i>National Center for Biotechnology Information</i>
ncRNA	Kodlamayan Ribonükleik Asit
NLP	<i>Natural Language Processing</i>
PDB	Protein Veri Bankası
RCSB	Protein Veri Bankası
RO5	<i>Lipinski's Rule of Five</i> (Beş Kuralı)
SQL	<i>Structured Query Language</i>
TPP	Tiamin prifosfat
TRUBA	Türk Ulusal Bilim e-Altyapısı
vb.	Ve Benzeri

ÖZET

Riboşalterlerin İnsan Genomunda Aranması ve Bilgisayar Ortamında Vaka Çalışmaları ile Potansiyel Bir İlaç Hedefi Olarak Tahmin Edilmesi

Riboşalterler genellikle mRNA'nın 5' ucu bölgelerinde bulunan genetik düzenleyici yapılardır. Riboşalterlerin en belirleyici özelliği, doğrudan bir fizyolojik sinyali tanımları ve RNA yapısında ortaya çıkan kaymalar ile spesifik genlerin regülasyonunu etkilemeleridir. Mevcut *Rfam* veri tabanı 40 riboşalter ailesi içermektedir. Riboşalterler, belirli efektörleri bağlamak için farklı moleküler belirleyiciler geliştirdiklerinden, sadece ideal geri bildirim mekanizmaları için değil, aynı zamanda yeni ilaç hedefleri olarak hizmet etme potansiyelleri için de ilgi toplamışlardır. Diğer taraftan, riboşalterlerin, geri besleme mekanizmaları yoluyla hücrede hayati yolların düzenlenmesinde önemli rolleri vardır. Bu özelliklerinden dolayı riboşalterler, çoklu ilaç direncine (MDR) bir çözüm bulmak amacıyla, teorik ve pratik olarak antibiyotik tasarımı için potansiyel yeni hedefler olarak kabul edilmiştir. Bu çalışmada iki amaç yer almaktadır. Birinci amacımız insan genomunda literatürde mevcut olan riboşalter yapılarıyla anlamlı derecede uyumluluk gösteren gen sekanslarının tespit edilmesi ve bu gen sekanslarının bilinmeyen fonksiyonlarının ortaya çıkarılmasıdır. Bu doğrultuda lizin (*lysine*) riboşalteri ile insan genomunun belirli bölgelerinde anlamlı kabul edilebilecek sınıra yakın eşleşmeler elde edilmiştir. Çalışmamızın ikinci amacı ise, “*Rfam*” biyolojik veri tabanında yer alan erişime açık verileri kullanarak, riboşalter RNA aile üyelerinin potansiyel bir ilaç hedefi olarak düşünüldüğü sanal tarama analizi süreçlerini organize eden bir ardışık düzen oluşturmaktır. Çalışma sonucunda oluşturduğumuz ardışık düzen “*V. vulnificus*” genomuna ait sanal tarama analizi sonuçlarını elde etmeyi başarmıştır. Elde edilen sanal tarama analizleri Lipinski kuralı dikkate alınarak incelendiğinde “*ZINC000040394717*” erişim kimliğine sahip molekülün “*Ceftazidime*”; “*ZINC000004097280*” erişim kimliğine sahip molekülün ise “*Ciprofloxacin*” ile benzer moleküller olduğu söylenebilir. *Ceftazidime* ve *Ciprofloxacin* “*V. vulnificus*” organizması kaynaklı hastalıkların tedavisinde kullanılan ilaçlar arasındadır. Benzerlik bulduğumuz ZINC erişim kimliğine sahip moleküllerin literatürde henüz bilinen bir kullanımı yoktur. Bu açıdan literatüre katkıda bulunulmuştur.

Anahtar Sözcükler: Ardışık düzen, *Rfam*, Riboşalter, Sanal tarama, ZINC.

ABSTRACT

Searching for Riboswitches in the Human Genome and Estimation of Them as A Potential Drug Target With Insilico Case Studies

Riboswitches are genetic regulatory structures usually located at the 5' end regions of mRNA. The most distinctive feature of riboswitches is that they directly recognize a physiological signal and affect the regulation of specific genes by shifts in RNA structure. The current release of Rfam database contains 40 families of riboswitches. As riboswitches have developed different molecular determinants to bind specific effectors, they have attracted attention not only for their ideal feedback mechanisms, but also for their potential to serve as new drug targets. On the other hand, riboswitches have important roles in regulating vital pathways in the cell through feedback mechanisms. Because of these properties, riboswitches have been considered as potential new targets for antibiotic design, theoretically and practically, in order to find a solution to multidrug resistance (MDR). This study has two purposes. Our primary aim is to identify gene sequences in the human genome that show significant compatibility with riboswitch structures available in the literature and to reveal the unknown functions of these gene sequences. In this direction, close to the boundary matches were obtained with the lysine riboswitch, which can be considered significant in certain regions of the human genome. The second aim of our study is to create a pipeline that organizes virtual screening analysis processes in which riboswitch RNA family members are considered as potential drug targets, using publicly available data from the "Rfam" biological database. The pipeline we created as a result of the study succeeded in obtaining the results of the virtual screening of the *V. vulnificus* genome. When the obtained virtual screening results are examined considering the *Lipinski* rule, It can be said that the molecule with the accession ID "ZINC000004097280" is similar to "Ciprofloxacin" and the molecule with the access ID "ZINC000040394717" is similar to "Ceftazidime". "Ceftazidime" and "Ciprofloxacin" are among the drugs used in the treatment of diseases caused by the *V. vulnificus* organism. Molecules with the ZINC accession identity that we have found similarity do not have a known use in the literature yet. In this respect, a contribution has been made to the literature.

Keywords: Pipe-line, Rfam, Riboswitch, Virtual Screening, ZINC.

1. GİRİŞ ve AMAÇ

Riboşalterler (*riboswitch*) genellikle mRNA'nın 5' ucu bölgelerinde bulunan genetik düzenleyici yapılardır (1–4). Proteinler için olduğu gibi riboşalterler için de tanımlanmış veri tabanları, tahmin etme ve üç boyutlu yapılarını modellemede kullanılan çeşitli yazılımlar bulunmaktadır. Örneğin “*Rfam*” (5, 6) veri tabanı (Kasım 2022, 4108 RNA ailesi) 40 riboşalter ailesi içermektedir. Riboşalterlerin en belirleyici özelliği, doğrudan bir fizyolojik sinyali tanımları ve RNA yapısında ortaya çıkan kaymalar ile spesifik genlerin regülasyonunu etkilemeleridir (7).

Riboşalterlerin çoğu, küçük moleküllü bir ligandı spesifik olarak bağlar. Ligand, genellikle bir biyosentetik yolun son ürünüdür ve bunun elde edilmesine veya sentezine katılan genler, ligand konsantrasyonu yüksek olduğunda, daha fazla biyosentetik kapasitenin gereksiz olduğu hücreye sinyal gönderildiğinde bastırılır. Daha az sıklıkla, riboşalterler regüle edilen yol için substrat görevi gören metabolitlere cevap verebilir ve substrat konsantrasyonu yüksek olduğunda yolağın ekspresyonu indüklenir. Metal iyonları kadar küçük moleküller de riboşalterler için ligand olarak işlev görebilirler. Bu da onların taşınması için yolları kapatır veya toksik etkileri önlemek için taşıyıcıları kodlayan genleri açarlar (8–12). Riboşalterler, belirli efektörleri bağlamak için farklı moleküler belirleyiciler geliştirdiklerinden, sadece ideal geri bildirim mekanizmaları için değil, aynı zamanda yeni ilaç hedefleri olarak hizmet etme potansiyelleri için de ilgi toplamışlardır. Bu duruma dair en önemli örnek, “*Riboflavin*” temel biyosentezini bakterilerde bloke eden doğal veya sentetik ligand analogları tarafından inhibe edilebilen “*Flavin Mononükleotid*” (*FMN*) riboşalteridir (13–17). Riboşalterlerin, protein-bağımsız bir şekilde hücre içi metabolitler için spesifik ve seçici bir sensör olarak işlev görme yeteneği, yeni tasarlanmış “*aptamerler*” ve biyosensörler geliştirmek için umut verici bir araştırma alanı olmasını sağlar. Diğer taraftan, riboşalterlerin, geri besleme mekanizmaları yoluyla hücrede hayati yolların düzenlenmesinde önemli rolleri vardır. Bu özelliklerinden dolayı riboşalterler, çoklu ilaç direncine (MDR) bir çözüm bulmak amacıyla, teorik ve pratik olarak antibiyotik tasarımı için potansiyel yeni hedefler olarak kabul edilmiştir (18).

Riboşalterler şimdiye kadar genellikle *Bacillus subtilis* (19, 20), *Vibrio vulnificus* (21), *Salmonella enterica serovar* (22) ve *Escherichia coli* (19) bakterilerinde tespit edilmiştir. Ökaryotlarda, sadece TPP bağlayıcı riboşalteri, gen ekspresyonunu durdurarak veya teşvik ederek ekleme işlemini kontrol edebileceği DNA bölgelerinde tanımlanmıştır

(19, 23). Bitkilerde, *Arabidopsis thaliana*, *Oryza sativa* ve *Poa secunda* (24, 25) gibi, THIC geninin 3' ucu bölgesi yüksek oranda korunmuştur ve bir TPP riboşalter barındırır. Bu tip mRNA'larda, başlangıç “*kodonunu*”; bir “*intron*”, küçük bir “*ekzon*” ve TPP riboşalterine sıkıca bağlanan ikinci bir “*intron*” izler (26). *Aspergillus oryzae* (27) ve *Neurospora crassa* (28) gibi mantarlarda, TPP riboşalteri 5' ucu bölgesinde yer alır. Bu organizmalarda TPP seviyeleri arttığında, riboşaltere bağlanan metabolit, “*intronunun*” bir kısmını muhafaza ederken alternatif bir ekleme bölgesini ortaya çıkarır (26). TPP riboşalteri, *Chlamydomonas reinhardtii* ve *Volvox carteri* gibi alglerden gelen THI4 ve THIC genlerinin transkripsiyonu esnasında benzer bir mekanizma kullanır (29).

Bu çalışmada iki amaç yer almaktadır. Birinci amacımız insan genomunda literatürde mevcut olan riboşalter yapılarıyla anlamlı derecede uyumluluk gösteren gen sekanslarının tespit edilmesi ve bu gen sekanslarının bilinmeyen fonksiyonlarının ortaya çıkarılmasıdır. Bu doğrultuda lizin (*lysine*) riboşalteri ile insan genomunun belirli bölgelerinde anlamlı olabilecek düzeyde eşleşmeler elde edilmiştir.

Çalışmamızın ikinci amacı ise, “*Rfam*” biyolojik veri tabanında yer alan erişime açık verileri kullanarak, riboşalter RNA aile üyelerinin potansiyel bir ilaç hedefi olarak düşünüldüğü sanal tarama analizi süreçlerini organize eden bir ardışık düzen oluşturmaktır.

Çalışma sonucunda oluşturduğumuz ardışık düzen “*V. vulnificus*” genomuna ait sanal tarama analizi sonuçlarını elde etmeyi başarmıştır. Her ne kadar sanal tarama analizi işlemleri için hazır bir program kullanmış olsak da tasarladığımız ardışık düzen içerisindeki organizasyonla alakalı tüm kod betikleri yazar tarafından oluşturulmuştur. Ayrıca oluşturduğumuz ardışık düzen herhangi bir “*Rfam*” üyesi RNA elemanına ait üç boyutlu yapı dosyası üzerinden de benzer sanal tarama analizini yapabilecek esnekliktedir.

2. GENEL BİLGİLER

2.1. Genel Bilgilerin Veriliş Şekli Hakkında Kısa Açıklama

Bu çalışma multidisipliner bir çalışmadır. Bu nedenle birçok farklı alandaki konularla ilgili kavramlara tez içerisinde yer verilmiştir. Genel bilgilerin verileceği bu kısımda yer alan içerikler ilk olarak ait oldukları disiplin alanı içerisinde tanımlanan alt başlıklar halinde verilmiştir. Ayrıca tez içerisinde Türkçe karşılığı olmayan çeşitli kavramlar ve ilgili alan terminolojisinde kalıplaşmış bazı kelimeler de kullanılmıştır. Bu kavramlar ve kelimeler orijinal yazım şekilleri ile italik düzende çift tırnak arasında verilmiştir. Genel bilgiler bölümünde tanımlandığı alt başlık içerisinde de açıklaması verilmiştir.

2.2. Biyoinformatik Veri Tabanları ve Bağlantılı Kavramlar Hakkında Bilgiler

2.2.1. Biyoinformatik Nedir?

Biyoinformatik, biyolojinin çeşitli alt dallarının yanı sıra moleküler biyoloji ve bilgisayar teknolojisinin sunduğu veri işleme aygıtlarını barındıran bilimsel disiplindir. Başka bir ifadeyle, karmaşık biyolojik verilerin düzenlenmesi ve analiz edilmesi işlemlerini gerçekleştiren bilim dalıdır (30).

2.2.2. Veri Tabanı Nedir?

Genel anlamıyla ifade etmek gerekirse veri tabanı düzenli bilgilerin tablo düzenindeki topluluğu demektir. Bilgisayar terminolojisinde açısından ifade etmek gerekirse, sistematik olarak erişilebilen, yönetilebilen, güncellenebilen, taşınabilen ve birbirleri arasında tanımlı ilişkiler bulunabilen veri kümesi anlamına gelir (31).

2.2.3. Biyolojik Veri Tabanı Nedir?

Mevcut olan biyolojik verilere araştırmacıların kolayca bulabilmeleri ve elde ettikleri yeni bulguları kolayca paylaşabilmeleri amacıyla bazı kuruluşlar internet ortamından ücretsiz erişilebilecek veri tabanları kurmuşlardır. Aynı zamanda bu kuruluşlar internet ortamında kullanılabilecek birçok biyoinformatik aracını da araştırmacıların hizmetine genellikle ücretsiz olarak sunmaktadırlar (32). Biyolojik veri tabanı örnekleri için (33) numaralı kaynakta bağlantısı verilen çevrimiçi tabloyu inceleyebilirsiniz. Bu tez çalışmasında kullanılan çeşitli biyolojik veri tabanları Tablo 1’de verilmiştir.

Tablo 1. Tez çalışmasında kullanılan çeşitli biyolojik veri tabanları

Veri Tabanı	Erişim Adresi
NCBI	https://www.ncbi.nlm.nih.gov/
Rcsb	https://www.rcsb.org/
RNA Central	https://rnacentral.org/
Pubmed	https://pubmed.ncbi.nlm.nih.gov/
Rfam	https://rfam.org/
ZINC	https://zinc.docking.org/
PubChem	https://pubchem.ncbi.nlm.nih.gov/

2.3. Rfam Biyoinformatik Veri Tabanı ve Bağlantılı Kavramları Hakkında Bilgiler

2.3.1. Rfam Nedir?

“Rfam”, kendi adıyla oluşturulmuş biyolojik veri tabanı verilerinin en son sürümünü içeren, herkese açık, salt okunur bir MySQL veri tabanına sahiptir. Bahsi geçen bu veri tabanı her yeni sürümü çıktığında güncellenmektedir (34). Rfam veri tabanı, kodlamayan RNA genlerinin yanı sıra *cis düzenleyici elemanları* içeren yapısal RNA'ların RNA dizisi ailelerinin bir koleksiyonudur. Her aile, bir *çoklu dizi hizalaması* ve bir *kovaryans modeli (CM)* ile temsil edilir (5, 6).

Rfam web sitesini (35) herhangi bir aile hakkında bilgi edinmek veya ailelere ve genom açıklamalarına göz atmak için kullanabilirsiniz. Ayrıca, tüm Rfam verilerini erişime açık dosya transfer sunucu (FTP) adresinden indirilebilir (36). Rfam, bünyesinde yer alan her bir RNA ailesi için Tablo 2’de verilmiş olan bilgileri içermektedir.

Tablo 2. Rfam üyesi her bir RNA elemanı için düzenlenen web sayfasına ait içerik türleri

İçerik adı	İçeriğe dair kısa açıklama
Summary page	Çevrimiçi ansiklopedi Wikipedia'dan elde edilen RNA ailesi hakkında metinsel arka plan bilgisi
Seed alignment	Küçük bir temsili dizi seti ve bir <i>konsensüs</i> ikincil yapı açıklamasını içeren küratörlü bir hizalama
Sequences	Bit puanı, tohum ve tam hizalamalar, bölge koordinatları, Avrupa Moleküler Biyoloji Laboratuvarı (EMBL) nükleotid veri tabanından dizi açıklaması ve tür adı dahil olmak üzere ailedeki diziler hakkında bilgiler
Secondary structure	Çeşitli sekans ölçümleri ve yapı koruma yöntemleri ile anotasyonlu (<i>annotated</i>) ikincil yapı görüntüleri
Species	Tam hizalama için tür dağılımını gösteren etkileşimli ağaç grafiği
Trees	Tohum (<i>seed</i>) ve tam hizalama için mevcut olan filogenetik ağaçlar
Structures	PDB yapıları ve Rfam açıklamaları arasındaki eşlemeler
Database references	Dış veri tabanlarına bağlantılar ve diğer veri kaynaklarına referanslar
Curation	Kovaryans model dosyaları, hizalamanın yazarı, dizi ve/veya yapı kaynakları için referanslar, her hizalamadaki dizi sayısı, puan eşikleri ve puan dağılımları dahil aileyi özetleyen bilgileri içerir.

2.3.2. Rfam Veri Tabanındaki RNA Aileleri Nasıl Oluşturuluyor?

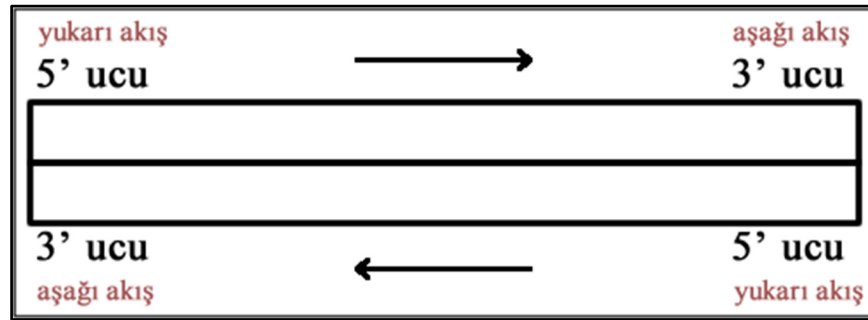
Aşağıdaki özelliklere sahip bir RNA elemanının varlığı, yeni bir aile oluşturmak için ideal ortamın oluştuğunu bize gösterir. Bu özellikler şunlardır:

- Bazı bilinen işlevsel sınıflandırmaya sahip olması
- Evrimsel olarak korunması
- Belirli bir ikincil yapısı olduğuna dair kanıtın var olması

Verilen özellikler doğrultusunda, yeni bir aile kurmak için, önce yayınlanmış literatürden deneysel olarak doğrulanmış en az bir örnek alınmalıdır. Literatürde başka homologlar tanımlanırsa, bunlarda tohum “*seed alignment*” verisine eklenmelidir. Alternatif olarak, bahsedilen içerikler mevcut değilse, benzerlik araştırması (“*Infernal*” kullanarak) veya manuel iyileştirme yoluyla ailenin diğer üyeleri belirlenmeye çalışılmalıdır (37).

2.3.3. Cis-Düzenleyici Eleman Nedir?

DNA'nın veya RNA'nın bir bölgesi olan Cis-düzenleyici eleman, genlerin ifade edilme şeklini düzenlemeye yarayan bir görev üstlenmektedir. Bulunduğu organizmaya ait kod dizilimi içerisinde, genlerin kodlanan bölgeleri içerisinde yukarı akış (*upstream*) veya aşağı akış (*downstream*) konumlarında bulunabileceği gibi kodlanmayan (*intron*) bölge konumlarında da bulunabilirler (38).



Şekil 1. DNA'daki aşağı akış ve yukarı akış bölgeleri

2.3.4 Kodlamayan RNA (non-coding RNA) Nedir?

Kodlamayan RNA, proteine dönüşümü yapılmayan fonksiyonel RNA moleküllerine denir. İngilizce kaynaklı literatürde “*non-coding RNA*” kısaltma olarak *ncRNA* olarak ifade edilir. Kodlamayan RNA'nın bulunduğu DNA sekans dizileri RNA geni ya da kodlamayan RNA geni olarak tanımlanır (39).

2.3.5. Klan (Clan) Nedir?

Bir *Rfam* klanı, ortak bir atayı paylaşan ancak makul bir şekilde hizalanamayacak kadar farklı olan veya hizalanabilen ancak farklı işlevleri olan bir aile grubudur. Örneğin, LSU klanı (CL00112), bakteriyel, ökaryotik ve *arkeal* LSU aileleri dahil olmak üzere farklı büyük ribozomal alt birim RNA tiplerini tanımlayan 5 aile içerir (40).

2.3.6. Aile (Family) nedir?

Sekans veya ikincil yapı bakımından evrimsel olarak ilişkili olduğuna inanılan bir RNA sekansı grubu (41).

2.3.7. Kovaryans Modeli (CM) Nedir?

RNA yapısal hizalaması için ikincil yapı profili olarak ifade edilir. *Rfam* kovaryans modelleri, *Infernal* yazılımı tarafından RNA veya DNA dizilerindeki RNA ailelerinin örneklerini bulmak için kullanılır (42).

2.3.8. Infernal Yazılımı Nedir?

Infernal, RNA yapısı ve dizi benzerlikleri için DNA dizisi veri tabanlarını aramak için oluşturulmuş bir yazılımdır. Kovaryans modelleri (CM) adı verilen özel bir profil stokastik bağlamdan bağımsız gramer durumunun bir uygulaması olarak çalışmaktadır (43).

2.3.9. Tohum Hizalama (Seed Alignment) Nedir?

Bir aile için manuel olarak seçilmiş bir temsili sekans dizisi örneğidir. Bu sekans dizileri hizalanır ve bir konsensüs ikincil yapısı ile açıklanır. Bu hizalama, aile için kovaryans modelini oluşturmak için kullanılır (44).

Seed sequence alignment for RF00078	
<i>Escherichia coli</i> str. K-12 substr. M51655	GAAAGACGGCA...UUUGUUAUCA...UCAUCGUAUU...CAG.AGA.UG...Aa.aUUUU.GGCCACU..CA-CG.....UGUGGCG..UUUU
<i>Baricosa sp. AT-9b</i>	GAAAGACGGCA...UUUGUUAUCA...UCAUCGUAUU...CAG.AGA.UG..A..A..UUUGGCCAC..GU-GA.....UGUGGCG..UUUU
<i>Versinia enterocolitica</i> subsp. enterocolitica 8081	GAAAGACGGCA...UUUGUUAUCA...UCAUCGUAUU...CAG.AGA.UG..A..A..UUUGGCCAC..GU-GA.....UGUGGCG..UUUU
<i>Bickeysa zeae</i> Ech1591	GAAAGACGGCA...UUUGUUAUCA...UCAUCGUAUU...CAG.AGA.UG..A..A..UUUGGCCAC..GU-GA.....UGUGGCG..UUUU
<i>Erwinia pyrifoliae</i> DSM 12163	GAAAGACGGCA...UUUGUUAUCA...UCAUCGUAUU...CAG.AGA.UG..A..A..UUUGGCCAC..GU-GA.....UGUGGCG..UUUU
<i>Salmonella enterica</i> subsp. enterica serovar Typhimurium str. SL1344	GAAAGACGGCA...UUUGUUAUCA...UCAUCGUAUU...CAG.AGA.UG..A..A..UUUGGCCAC..GU-GA.....UGUGGCG..UUUU
<i>Xenorhabdus nematophila</i> ATCC 19061	G-AAGAGCGCAaaa-UUGUUAU-...U-AUCGUAUUU...CUA-.AGA.UAGUUU..A..-UUUGGCCAC..-UCC-.....UGUGGCG..UUUU
<i>Sodalis glossinidius</i> str. "morsitans"	GAAAGACGGCA...UUUGUUAUCA...UCAUCGUAUU...CAG-.GAUG.UU-..A..AUUUA-GCCAC..GUUC-.....UGUGGCG..UUUU
<i>Arsenophonus nasoniae</i> (non-killer infecting Nasonia vitripennis)	GCAAGACGGC-Aaaa-UUGUUAU-...-DAUC-UGUUCU...CUA-.AGA-GAUUU.UA..UUUU.GGCCACU..UUA-.....AGUGGCG..UUUU
<i>Enterobacter cloacae</i> SCF1	GAAAGACGGCA...UUUGUUAUCA...UCAUCGUAUU...CAG.AGA.UG...-AaaUGUUU.GGCCAC..GU-GA.....UGUGGCG..UUUU
<i>Xenorhabdus bovienii</i> SS-2004	G-AAGAGCGCAaaa-UUGUUAU-...-CAUCGUAUUU...CUA-.AGA-GaUUU.UA..-UUUGGCCAC..ACU-.....UGUGGCG..UUUU
<i>Riesbia pneumoniae</i> KCTC 2242	GAAAGACGGCA...UUUGUUAUCA...CAUCGUAUUU...CAG.AGA.UG...-Aa.aUUUU.GGCCAC..GU-GA.....UGUGGCG..UUUU
<i>Bickeysa dadantii</i> 3927	GAAAGACGGCA...UUUGUUAUCA...UGAUCGUAUUU...-CAG.AGA.UG.UUU...-CUUU.GGCCAC..GAGGACaUUGaUGUGGCG..UUUU
<i>Providencia rustigianii</i> DSM 4541	GAAAGACGGCAaaa-UUGUUAU-...-CAUCGUAUUU...CUAAG.AGA.U-..UUUUA..-UUU.GGCCAC..UUAU...-UGUGGCG..UUUU
<i>Candidatus Regiella insecticola</i>	GCAAGACGGCA...UUUGUUAU-...-CAUCGUAUUU...AUA.C.AGA.UG.UUU..-c..UUUU.GGCCAC..UA-AU...-AGUGGCG..UUUU
<i>Serratia sp.</i>	GAAAGACGGCA...UUUGUUAUCA...UCAUCGUAUUU...CAG.AGA.UG..A..A..UUUGGCCAC..GU-GA.....UGUGGCG..UUUU
<i>Proteus mirabilis</i>	G-AAGAGCGCAaaa-UUGUUAU-...-CAUCGUAUUU...CUA-.AGA-..AUUU..A..-UUUGGCCAC..CAAAUUU..UGUGGCG..UUUU

Resim 1. İkincil yapı sarmal bölgeleriyle renklendirilmiş örnek bir tohum hizalaması

2.3.10. Tam Hizalama (*Full Alignment*) Nedir?

Belirli bir *Rfam* ailesinin kovaryans modeli için manuel olarak ayarlanan eşik değerlerinden daha yüksek puan alan ilgili diziler kümesinin hizalanmasıdır (45).

2.3.11. Aileye Dahil Etme Eşik Değeri (*Gathering Cutoff*) Nedir?

Aileyi oluştururken *Rfam* küratörleri tarafından belirlenen bit puanı toplama eşiğidir. Bu eşikte veya üzerinde puan alan tüm diziler, tam hizalamaya dahil edilecek ve modelin gerçek homologları olduklarına inanılacaktır (46).

2.4. Biyoinformatik Alanına Yönelik Temel Kavramlar

2.4.1. RNA ve mRNA Nedir?

RNA, nükleotitlerin birleşiminden oluşan bir polimer yapıdır. RNA önemli biyolojik işlevlerin çoğunun gerçekleşmesinde görev alır, bu işlevlerin içerisinde DNA'da taşınan genetik kodun proteine dönüşümü (translasyon) ile ilişkili çeşitli görevler de bulunmaktadır.

RNA türlerinden biri olan mesajcı RNA (mRNA) protein sentezinin gerçekleşmesi için, DNA'daki ilgili kodu ribozomlara taşır (47).

2.4.2. BLAST ve FASTA Nedir?

BLAST, *GenBank* veri tabanı içerisinde yer alan nükleik asit ve protein sekans dizilerinin birbirlerine olan benzerliklerinin tespit edilmesi için tasarlanmış, biyoinformatik çalışmalarında kullanılan bir çeşit arama motorudur. İçerisinde barındırdığı tüm sekanslarla, kullanıcı tarafından girilen veya dizisi bilinen ama tanımlı olmayan gen dizilerini karşılaştırır. Bu karşılaştırma sonucunda benzerlik oranı kriterini dikkate alarak eşleşme bulunduğu gen bölgelerini yüksekte aşağıya doğru olacak şekilde listeler.

Bu genler listelenirken *GenBank* erişim numaraları ile gösterilir. Bunlara ait protein ve nükleik asit sekansları ise FASTA formatında gösterilir ve analizler bu format kullanılarak yapılır (48).

2.4.3. GFF Dosyası Nedir?

Genel özellik formatı (gen bulma formatı, jenerik özellik formatı, *GFF*), DNA, RNA ve protein dizilerinin genlerini ve diğer özelliklerini tanımlamak için kullanılan bir dosya formatıdır. Tüm GFF formatları (GFF2, GFF3 ve GTF), her satırda sekmeye ayrılmış dokuz

alandan oluşmaktadır. Tüm GFF formatları ilk yedi alan için aynı yapıyı paylaşırken, dokuzuncu alanın içeriği ve biçimi farklıdır (49).

2.4.4. PDB ve Idx Dosyası Nedir?

PDB dosyası, proteinler ve nükleik asitler gibi büyük biyolojik moleküllerin üç boyutlu yapısal verilerini içeren bir dosyadır. Orijinal dosya formatı, bilgisayar delikli kartlarının genişliği nedeniyle satır başına 80 karakterle sınırlandırılmıştı. 1996 civarında, CIF formatının bir uzantısı olan "*makromoleküler Kristallografik Bilgi dosyası*" formatı, *mmCIF* aşamalı olarak devreye alındı. *mmCIF*, 2014 yılında PDB arşivi için standart format oldu (50).

IDX dosya formatı, çeşitli sayısal türlerdeki vektörler ve çok boyutlu matrisler için basit bir formattır (51). Veri tabanlarında hızlı arama yapmak için de kullanılır.

2.4.5. İntron ve Ekzon Nedir?

Protein ve enzimler elde edilirken, DNA üzerindeki genlerin kod dizilimi dikkate alınarak bu genlere denk gelen RNA dizilimlerinin kopyası oluşturulur. Bu RNA'ların (mRNA-mesajcı RNA) sentezlenme işleminden sonra, genin bazı kısımları Uç Birleştirme (*RNA Splicing*) adı verilen bir süreç ile RNA'nın içinden atılır ve geride kalan kısımlarda (*exonlar*) birleştirme yapılır.

DNA'nın okunduktan sonra atılan bu kısma *intron* denir. *Intronlar*, mRNA ve protein kodlaması süreçlerinde görev almazlar.

Genlerin kodlamada yer almayan bu kısmı, insan genomunun tamamının yaklaşık %97'lik bir bölümünü oluşturmaktadır. Literatürde, kodlanan kısımlar için ise ekzon (*exon*) adı kullanılmaktadır (52).

2.4.6. Gap, Gap Penalty ve E-değeri Nedir?

Boşluk cezası (*gap penalty*), iki veya daha fazla dizinin hizalamalarını puanlama yöntemidir. Dizileri hizalarken, dizilere boşluklar (*gaps*) eklemek, bir hizalama algoritmasının boşluksuz bir hizalamaya kıyasla daha fazla terim eşleştirmesine izin verebilir (53).

BLAST E-değeri, tesadüfen bulunabilecek benzer kalitede (puan) beklenen isabetlerin sayısıdır. E-değeri ne kadar küçük olursa eşleşme o kadar iyi olur (54).

2.4.7. GC Oranı Nedir?

Moleküler biyoloji ve genetikte, GC oranı (veya guanin-sitozin oranı), bir DNA veya RNA molekülündeki G veya C olan azotlu bazların yüzdesidir (55).

2.4.8. Annotation Nedir?

Genom anotasyonu (*annotation*), bir genom sekans dizisi boyunca fonksiyonel öğeleri tanımlama ve böylece ona anlam verme işlemidir. DNA dizilimi bilinmeyen işlevli birçok sekans dizisi ürettiği için bu işlem gereklidir (56).

2.4.9. Pairwise Alignment Nedir?

İkili sekans hizalaması, iki biyolojik sekans (protein veya nükleik asit) arasındaki işlevsel, yapısal ve/veya evrimsel ilişkileri gösterebilen benzerlik bölgelerini belirlemek için kullanılır (57).

2.4.10. Secondary Structure Prediction Nedir?

İkincil yapı tahmini, proteinlerin ve nükleik asit dizilerinin ikincil yapılarını yalnızca birincil yapılarının bilgisine dayanarak tahmin etmeyi amaçlayan biyoinformatikteki bir dizi tekniktir. Proteinler için bu, alfa sarmalları ve beta iplikçikleri gibi protein yapılarının oluşumunu tahmin etmek anlamına gelirken, nükleik asitler için baz eşleştirme ve baz istifleme etkileşimleri yoluyla sarmallar ve kök-halka yapıları gibi nükleik asit yapılarının oluşumunu tahmin etmek anlamına gelir (58).

2.4.11. Sunburst Grafiği Nedir?

“Sunburst” grafiği, tipik olarak hiyerarşik veri yapılarını görselleştirmek için kullanılır. Halka grafik, çok düzeyli pasta grafik ve radyal ağaç haritası olarak da bilinir. Bir “sunburst” grafiği, daha derin hiyerarşi seviyelerine sahip halkalarla çevrelenmiş bir iç çemberden oluşur. Her parçanın açısı ya bir değerle orantılıdır ya da üst düğümü altında eşit olarak bölünür. “Sunburst” grafiğindeki tüm bölümler, ait oldukları kategori veya hiyerarşi seviyesine göre renklendirilebilir (59).

2.5. Bilgisayar ve Kodlama Terminolojisiyle İlgili Kavramlar

2.5.1. Phyton ve PyCharm Nedir?

Python, nesne yönelimli, dinamik semantik ile harmanlanmış, ileri seviye bir programlama dilidir. Dinamik yazım desteği ve dinamik bağlama desteği ile birleştirilmiş

ileri düzeyde yerleşik veri yapıları özelliğine sahiptir. Bu özelliği sayesinde kullanıcılar onu, hızlı uygulama geliştirmek için tercih etmektedir. *Python*'un oldukça basit ve öğrenmesi kolay bir sözdizimi vardır. Bu sayede okunabilir kod geliştirme konusunda kullanıcıya destek olur ve program güncelleme maliyetini düşürür. Python, program işlevselliğini ve kodun yeniden kullanılabilmesine olanak veren modülleri ve paketleri barındırır ve yenilerini de destekler. Python derleyicisi ve kapsamlı standart kütüphanesi, ücretsiz olarak temin edilebilir (60).

PyCharm, çapraz platform (Windows, OS X ve GNU/Linux işletim sistemleri) üzerinden çalışan Python geliştirme ortamlarından bir tanesidir. *PyCharm Community Edition*, erişime açık olan ve ücretsiz olarak temin edilebilen bir *PyCharm* sürümüdür. Kullanımı, yaygın olarak öğrenciler tarafından yapılmaktadır. Ücretli olan ve içerisinde gelişmiş özellikleri barındıran sürümü ise *Professional Edition* olarak adlandırılmaktadır (61).

2.5.2. *Anaconda* ve *PyPI* Nedir?

Python ve R programlama dillerine yönelik olarak hazırlanmış paketlerin bir dağıtıcısı olan *Anaconda*, paketler (veri bilimi, makine öğrenmesi uygulamaları, büyük boyutlu veri işleme, tahmine dayalı analiz vb.) ile ilgili süreçleri basitleştirmek amacıyla oluşturulmuştur. *Anaconda*, çapraz platforma (Windows, OS X ve GNU/Linux işletim sistemleri için) uygun çeşitli veri bilimi paketlerini içerir. 2012 yılında Peter Wang ve Travis Oliphant tarafından kurulan “*Anaconda*” şirketi tarafından geliştirilmiş ve sürdürülmüştür (62).

PyPI olarak kısaltılan Python Paket Dizini, *Python* için resmi üçüncü parti yazılım deposudur (63).

2.5.3. *SQL* ve *MYSQL* Nedir?

SQL olarak kısaltılan “Yapılandırılmış Sorgu Dili”, programlamada kullanılan ve ilişkisel veri tabanı yönetim sisteminde (RDBMS) tutulan verileri yönetmek veya ilişkisel veri akışı yönetim sisteminde (RDSMS) akış işlemek için tasarlanmış alana özgü bir dildir. Özellikle, yapılandırılmış verilerin, yani varlıklar ve değişkenler arasındaki ilişkileri içeren verilerin işlenmesinde yararlıdır (64).

2.5.4. Web Servisi ve Api Nedir?

Bir bilgisayar aygıtında çalışan, bir ağ üzerinden belirli bir bağlantı noktasındaki istekleri dinleyen ve karşılığında yanıt olarak web belgesi sunan bir sunucuya web servisi denir (65).

Bir uygulama programlama arayüzü (API), iki veya daha fazla bilgisayar programının birbiriyle iletişim kurmasının bir yoludur. Diğer yazılım parçalarına hizmet sunan bir tür yazılım arayüzüdür (66).

2.5.5. Pipe-Line (Ardışık Düzen) Nedir?

Pipe-line (ardışık düzen), bir veya daha fazla veri setinin bir dizi kronolojik adımla değiştirildiği bir bilgi işlem uygulamasıdır. Adımlar tipik olarak ardışıktır ve her biri bir sonrakini veri setinin değiştirilmiş versiyonuyla besler. Veriler tüm adımlardan geçtikten sonra işlem hattı tamamlanır ve elde edilen veriler daha sonra yeni bir hedefe yazılır veya gönderilir (67).

2.5.6. Regular Expression (Düzenli İfade) Nedir?

Düzenli ifade, metinde bir arama modeli belirten bir karakter dizisidir. Genellikle bu tür modeller, dizi arama algoritmaları tarafından dizelerde "bul" veya "bul ve değiştir" işlemleri veya girdi doğrulama için kullanılır. Düzenli ifade teknikleri, teorik bilgisayar bilimi ve biçimsel dil teorisinde geliştirilmiştir (68).

2.5.7. TRUBA ve Slurm Nedir?

TÜBİTAK ULAKBİM, ülkemizde bulunan tüm araştırma kurumları ile araştırmacılara yüksek başarımlı hesaplama yapabilme ve veri depolayabilme imkânı sağlayan ulusal bir merkezdir. Açık ifadesi Yüksek Başarımlı ve Grid Hesaplama Merkezi olan bu merkez, TÜBİTAK Bilim Kurulu kararı ile kısaltması TRUBA olan Türk Ulusal Bilim e-Altyapısı olarak da anılmaktadır. Araştırmacıların çeşitli ortamlarda diğer ülkelerdeki araştırmacılar ile denk koşullarda çalışabilmesi için gerekli olan işlem, analiz ve veri saklama imkânlarının, genişletilerek güçlendirilmesi, Türkiye Bilim e-Altyapısı alanına yönelik kurulan Danışma Kurulu'nun öneri ve görüşleri doğrultusunda planlanmaktadır (69).

Slurm, büyük ve küçük Linux kümeleri için açık kaynaklı, hataya dayanıklı ve yüksek düzeyde ölçeklenebilir bir küme yönetimi ve iş planlama sistemidir. *Slurm*, çalışması

için hiçbir çekirdek modifikasyonu gerektirmez ve nispeten bağımsızdır. Bir küme iş yükü yöneticisi olarak *Slurm*'un üç temel işlevi vardır. İlk olarak, kullanıcılara belirli bir süre için kaynaklara (bilgi işlem düğümleri) özel ve/veya özel olmayan erişim tahsis eder, böylece iş yapabilirler. İkinci olarak, tahsis edilen düğümler kümesinde işin (normalde paralel bir iş) başlatılması, yürütülmesi ve izlenmesi için bir çerçeve sağlar. Son olarak, bekleyen bir iş kuyruğunu yöneterek kaynakların dağıtımına hakemlik eder (70).

2.5.8. Çalışmada Kullanılan Çeşitli Dosya Türleri Hakkında Açıklamalar

Bu çalışma için oluşturulan ardışık düzen içerisinde birçok farklı türde dosya kullanılmıştır. Tablo 3'te bu dosya türleri hakkında kısa açıklamalar verilmiştir.

Tablo 3. Çalışmada kullanılan dosya türleri ve açıklamaları

Dosya türü	Dosya türüne dair kısa açıklama
<i>xml</i>	Yapılandırılmış bilgileri temsil etmek için basit bir metin tabanlı dosya biçimi
<i>html</i>	Web sayfalarını oluşturmak için kullanılan standart metin işaretleme dili dosya biçimi
<i>json</i>	Küçük boyutlu veri değişim dosya biçimi
<i>csv</i>	Virgülle ayrılmış değerler dosya biçimi
<i>log</i>	Günlük dosyası, bir işletim sistemi, uygulama, sunucu veya başka bir aygıt içindeki kullanım kalıpları, etkinlikler ve işlemler hakkında bilgiler içeren, bilgisayar tarafından oluşturulmuş bir veri dosya biçimi
<i>svg</i>	Web için vektör tabanlı grafikleri tanımlayan dosya biçimi
<i>png</i>	Hemen hemen her grafik programında görüntüleyebileceğiniz yüksek kaliteli bir grafik dosyası biçimi
<i>tar.gz</i>	Daha çok Linux tabanlı bilgisayarlarda kullanılan sıkıştırılmış dosya biçimi

2.6. Sanal Tarama Analiziyle İlgili Kavramlar

2.6.1. Ligand Nedir?

Merkezi bir metal atomuna veya iyonuna bağlanan iyonlara ya da yüksüz olan moleküllere ligand adı verilir. Ligandlar en az bir donör atoma sahiptir. Bu donör atomu sahipliği, merkez atomla kovalent bağı oluşturabilmek için kullanılır (71). Bu çalışmada “*ligand*” terimi yerine ilaç adayı molekül terimi kullanılacaktır.

2.6.2. Reseptör Nedir?

Biyokimya ve farmakolojide reseptörler, biyolojik sistemlere entegre edilebilecek sinyalleri alan ve ileten, proteinden oluşan kimyasal yapılardır. Belirli bir tipteki reseptörler, sinyale karşılık gelen spesifik hücresel biyokimyasal yollara bağlıdır. Çoğu hücrede çok sayıda reseptör bulunurken, her reseptör yalnızca belirli bir yapıdaki ligandlara bağlanır (72).

2.6.3. Docking (Moleküler Yanařtırma) Nedir?

Moleküler modelleme disiplinde, yanařtırma (*docking*), bir ligandın bir makromolek l ile kararlı bir kompleks oluřturma amacıyla birbirine baėlandıėında tercih edeceėi yerleřimi tahmin etmeye  alıřan bir y ntem olarak ifade edilir. Tercih edilen yerleřim bilgisi iki molek l (ligand ve makromolek l) arasındaki birleřme veya baėlanma enerjisini tahmin etmek i in kullanılabilir (73).

2.6.4. Blind Docking Nedir?

Molek ler k r yanařtırma, hedef cep hakkında herhangi bir  n bilgi olmaksızın bir ligandın bir proteinin t m y zeyine kenetlenmesi anlamına gelir. K r kenetleme, uygun bir protein-ligand kompleksi pozu bulunmadan  nce birka  deneme/ alıřma ve birka  enerji hesaplaması i erir (74).

2.6.5. AD4 Nedir?

AutoDock (AD), bir otomatik molek ler yanařtırma ara ları paketidir. Substratlar veya ila  adayları gibi k  k molek llerin bilinen    boyutlu yapıya sahip bir resept re nasıl baėlanacaėını tahmin etmek i in tasarlanmıřtır. Yıllar ge tik e, yeni iřlevler eklemek i in deėiřtirildi ve geliřtirildi. *AutoDock*'un mevcut daėıtımları, iki t r yazılımdan oluřur: *AutoDock 4* (AD4) ve *AutoDock Vina* (75).

2.6.6. AutoDock Vina Nedir?

AutoDock Vina, molek ler yanařtırma yapmak i in kullanılan a ık kaynaklı bir programdır. *Vina*'nın tasarım felsefesi, kullanıcının uygulama ayrıntılarını anlamasını, belirsiz arama parametrelerini deėiřtirmesini, sonu ları k melemesini veya geliřmiř cebir (*d rdeyle*) bilmesini gerektirmez. Gereken tek řey, kenetlenmekte olan molek llerin yapıları ve baėlanma b lgesi dahil arama uzayını belirlemektir. Izgara haritalarının hesaplanması ve atom y klerinin atanması gerekli deėildir (76, 77).

2.6.7. BFGS Algoritması Nedir?

Sayısal optimizasyonda, *Broyden–Fletcher–Goldfarb–Shanno* (BFGS) algoritması, kısıtlanmamıř doėrusal olmayan optimizasyon problemlerini   zmek i in yinelemeli bir y ntemdir. Davidon–Fletcher–Powell y nteminde olduėu gibi, BFGS de eėimi eėrilik bilgisi ile  nceden kořullandırarak iniř y n n  belirler (78).

2.6.8. *Exhaustiveness* Nedir ve Neden Önemlidir?

Vina'da, yerleştirme hesaplaması, rastgele yapılardan başlayan bir dizi bağımsız çalıştırmadan oluşur. Bu çalıştırmaların her biri bir dizi sıralı adımdan oluşur. Her adım, konformasyonun rasgele bozulmasını ve ardından yerel bir optimizasyonu (BFGS algoritmasını kullanarak) ve adımın kabul edilip edilmediği bir seçimi içerir. Her bir yerel optimizasyon, puanlama fonksiyonunun ve bunun konum-yönlendirme-burulma koordinatlarındaki türevlerinin birçok değerlendirmesini içerir. Bir yerel optimizasyondaki değerlendirme sayısı, yakınsama ve diğer kriterler tarafından yönlendirilir. Bir çalışmadaki adımların sayısı, ligandın ve esnek yan zincirlerin boyutuna ve esnekliğine bağlı olarak buluşsal olarak belirlenir. Ancak bunlardan kaç tanesinin çalıştırılacağına dair sayı *exhaustiveness* (ayrıntılılık) parametresi tarafından belirlenir (77).

2.6.9. *PyMOL*, *PDBQT* ve *MOL2* Nedir?

PyMOL, üç boyutlu moleküler yapıları işlemek ve görüntülemek için güçlü ve kapsamlı bir moleküler görselleştirme yazılımıdır (79).

PDBQT olarak adlandırılan genişletilmiş bir PDB formatı, atomik kısmi yükleri ve atom türlerini içeren koordinat dosyaları için kullanılır (80).

MOL2, tekli veya çoklu kimyasal bileşikler temsil eden ve bu bileşiklere ait her bir molekülün atomik koordinatlarını, kimyasal bağ bilgilerini ve meta verilerini tutan düz metin tablo formatıdır (81).

2.6.10. *Virtual Screening* (VS) Nedir?

Sanal tarama (VS), bir ilaç hedefine, tipik olarak bir protein reseptörüne veya enzime bağlanma olasılığı en yüksek olan yapıları belirlemek amacıyla küçük moleküllerin kitaplıklarını aramak için ilaç keşfinde kullanılan bir hesaplama tekniğidir (82).

2.6.11. *Lipinski'nin Beş Kuralı* Nedir?

Kısaca beş kuralı (RO5) olarak da bilinen *Lipinski'nin* beş kuralı, bir kimyasal bileşiğin ilaç adayı olup olamayacağını değerlendirmek, belirli bir farmakolojik veya biyolojik aktiviteye sahip bir kimyasal bileşiğin kimyasal özelliklere ve fiziksel özelliklere sahip olup olmadığını belirlemek için kullanılan pratik bir kuraldır. Kural, 1997'de *Christopher A. Lipinski* tarafından, ağızdan uygulanan ilaçların çoğunun nispeten küçük ve orta derecede lipofilik moleküller olduğu gözlemine dayanarak formüle edilmiştir (83, 84).

2.6.12. Tanimoto Benzerliđi Nedir?

Kimyasal benzerlik (veya moleküler benzerlik), kimyasal elementlerin, moleküllerin veya kimyasal bileşiklerin yapısal veya fonksiyonel niteliklere göre benzerliğini, yani kimyasal bileşimin inorganik veya biyolojik ortamlarda reaksiyon ortakları üzerindeki etkisini ifade eder. Parmak izleriyle temsil edilen kimyasal yapıları karşılaştırmak için en popüler benzerlik ölçüsü, Tanimoto (veya Jaccard) katsayısı T 'dir. $T > 0.85$ ise iki yapı genellikle benzer kabul edilir (85).

2.7. İstatistik Analizleriyle İlgili Kavramlar

2.7.1. Tek Örneklem T-Testi Nedir?

Tek örneklem t testi, bir örneklemin ortalamasını bilinen bir skorla (veya popülasyon ortalamasıyla) karşılaştırır. Test, bilinen bir popülasyon standart sapmasını veya bir örnek standart sapmasını kullanır. Test, aralık puanlarını analiz eder (86).

2.7.2. NLP (Doğal Dil İşleme) Nedir?

Bir dil, bilgi iletmek veya bilgiyi yayınlamak için sembollerin birleştirildiđi ve kullanıldığı bir kurallar veya semboller seti olarak tanımlanabilir. Tüm kullanıcılar makineye özgü dil konusunda bilgili olamayabileceğinden, Doğal Dil İşleme (NLP), yeni dilleri öğrenmek veya uzmanlık elde etmek için yeterli zamanı olmayan kullanıcılara hitap eder. Aslında, NLP, bilgisayarların insan dillerinde yazılmış ifadeleri veya kelimeleri anlamasını sağlamaya adanmış bir yapay zekâ ve dilbilgisi kitabıdır (87).

2.7.3. LDA (Doğrusal Ayırma Analizi) Nedir?

Kısaca ifade etmek gerekirse, Doğrusal Ayırma Analizi (LDA), Fisher'in doğrusal ayırma (*linear discriminant*) yönteminin genelleştirilmiş versiyonudur. Detaylı ifade etmek gerekirse LDA, en az iki nesne ya da olay sınıfını tespit eden veya ayıran özelliklerin doğrusal bir kombinasyonunun bulunması amacıyla başta istatistik alanı olmak üzere çeşitli diğer alanlarda da kullanımı tercih edilen bir yöntemdir. Ortaya çıkan kombinasyon, doğrusal bir sınıflandırıcı olarak kullanılabileceđi gibi, genellikle sonraki sınıflandırmadan önce boyut indirgeme işlemi için kullanılır.

LDA, aynı zamanda varyans analizi (ANOVA) ve regresyon analizi ile yakından ilişkilidir (88).

2.7.4. Wordcloud (Kelime Bulutu) Nedir?

Bir kelime bulutu (görsel tasarımda etiket bulutu veya ağırlıklı liste olarak da bilinir), genellikle web sitelerinde anahtar kelime meta verilerini tasvir etmek veya serbest biçimli metni görselleştirmek için kullanılan metin verilerinin görsel bir temsidir. Etiketler genellikle tek sözcüklerdir ve her etiketin önemi yazı tipi boyutu veya rengiyle gösterilir (89).

2.7.5. Text Mining (Metin Madenciliği) Nedir?

Metin madenciliği, metinden yüksek kaliteli bilgi elde etme işlemidir. Farklı yazılı kaynaklardan otomatik olarak bilgi çıkararak yeni, önceden bilinmeyen bilgilerin bilgisayar tarafından keşfedilmesini içerir. Yazılı kaynaklar arasında web siteleri, kitaplar, e-postalar, incelemeler ve makaleler olabilir. Yüksek kaliteli bilgi, tipik olarak istatistiksel model öğrenme gibi yollarla modeller ve eğilimler tasarlanarak elde edilir (90).

2.7.6. ICC (Sınıf İçi Korelasyon Katsayısı) Nedir?

İstatistik biliminde, sınıf içi korelasyon ya da diğer bir adıyla sınıf içi korelasyon katsayısı (ICC), gruplar halinde oluşturulmuş birimler üzerinde nicel veri ölçümleri yapıldığında kullanılabilen bir tanımlayıcı istatistik yöntemidir. ICC analizi, aynı gruptaki birimlerin birbirine olan benzerlik seviyesini tespit eder. Bir korelasyon türü gibi düşünülse de diğer birçok korelasyon ölçüsünün aksine, eşleştirilmiş gözlemler olarak yapılandırılmış veri türleri yerine gruplar olarak yapılandırılmış veri türleri üzerinde çalışır. Diğer bir tanımıyla ICC, aynı veriyi ölçen farklı gözlemciler tarafından yapılan kantitatif ölçümlerin tutarlılığının veya tekrarlanabilirliğinin değerlendirilmesidir (91).

3. GEREÇ VE YÖNTEM

Bu tez çalışmasında erişime açık çeşitli biyolojik veri tabanlarından derlenen veriler kullanılmıştır. Derlenmiş olan bu veriler kullanılarak ribozalter adlı RNA ailesine ait üç boyutlu yapılar elde edilmiştir. Elde edilen üç boyutlu yapılardan *V. vulnificus* bakterisine ait olan yapı kullanılarak bu bakterinin neden olduğu sağlık sorunlarını tedavi etmede kullanılabilecek potansiyel bir ilaç molekülü sanal tarama analizi ile aranmıştır. Sanal tarama analizi için gerekli olan aday ilaç molekülleri de erişime açık olan “ZINC” veri tabanına ait statik dosyalardan çekilmiştir. Analiz sonucunda bulunan en iyi sonuçlar ve bu sonuçların kaynağı olan aday ilaç molekülleri raporlanmıştır. Bununla birlikte tez çalışma konusunu oluşturan ribozalter adlı RNA ailesine ait “Rfam” veri tabanının çeşitli biyolojik veri tabanlarından derlediği veri ve içeriklerin çalışmamız kapsamında olanları da kendi geliştirdiğimiz ardışık düzen kod betikleri ile elde edilmiştir. Mevcut veri ve içeriklere çeşitli ilaveler yapılarak veri zenginleştirilmesi de yapılmıştır.

İnsan genomunda ribozalter benzeri yapıların olup olmadığına dair analizler de bu tez çalışmasının kapsamında yer almaktaydı. Bu soruya cevap alabilmek için “BLAST” analizi ile olası sekans eşleşmeleri tespit edilmiştir. Ardından “Infernal” yazılımı ile ikili sekans hizalaması (*pairwise alignment*) ve ikincil yapı tahmini (*secondary structure prediction*) analizleri de yapılmıştır. Bu ileri analizlerin yapılmasındaki amaç elde edilen sekans eşleşmelerinin anotasyonu (*annotation*) olarak verilecek boyutta olup olmadığının tespit edilebilmesidir. Tez çalışmasında bahsi geçen işlemler için “Phyton” programlama dili kullanılmıştır.

3.1. Veri Seti Derleme Prosedürü Hakkında Açıklama

Bu çalışmanın veri seti derleme sürecinin merkezinde “Rfam” veri tabanı yer almaktadır. “Rfam”, kendi web sitesinde yer alan tüm içerikleri barındıran veri tabanını, erişime açık *SQL* sunucusu üzerinden araştırmacıların erişimine sunmaktadır. Biz de çalışmamızda kullandığımız veri setlerinin bir kısmına bahsi geçen *SQL* sunucusu üzerinden ulaştık. Sanal tarama analizi için gerekli olan ilaç moleküllerine ise yine erişime açık olarak hizmet veren “ZINC” veri tabanı üzerinden ulaştık. Bunun yanında ilave analizler ve çeşitli veri zenginleştirme çalışmaları için de *Rcsb*, *Pubmed*, *Rna Central*, *PubChem* ve *Ncbi* biyolojik veri tabanlarından da veri derleme işlemi yaptık.

3.2. Geliştirilen Ardışık Düzen Hakkında Ön Bilgilendirme

Bu çalışma kapsamında geliştirilen ardışık düzen “*Phyton*” programlama diline ait kod betiklerinden oluşmaktadır. Ardışık düzen mantığına uygun olacak şekilde her bir kod betiğinin ürettiği çıktılar kendisinden sonraki kod betiği için gerekli girdileri oluşturmaktadır. Çalışma için geliştirilen ardışık düzen için, yazar tarafından oluşturulan kod betikleri üç ana başlık altında toplanmıştır. Bunlar “*rfam_part*” (Rfam veri tabanı kısmı), “*zinc_part*” (ZINC veri tabanı kısmı) ve “*docking_part*” (Sana tarama analizi kısmı) olarak her biri ayrı birer klasör olarak oluşturulmuştur. Her bir kod betiği Tablo 4’te verilen şekilde ait olduğu başlığın klasör konumu altında yazılmıştır.

Tablo 4. Tez çalışması için geliştirilen ardışık düzen süreçlerine ait kod betikleri

Süreç Başlığı	Kullanılan Kod Betiği
<i>rfam_part</i> (Rfam veri tabanı kısmı)	0_rFamDbBuilder.py
	1_rFamFamilyMembersDetector.py
	2_rFamAlignmentDataBuilder.py
	3_rFamFASTADownloader.py
	4_rFamGCRatioAnalyser.py
	5_rFamBlaster.py
	6_rFamBlastResultsAnnotator.py
	7_rFamHumanGenomeAnnotationsDescriptiver.py
	8_rFamTaxonomicDataPuller.py
	9_rFamTaxonomicDataPlotter.py
	10_rFamLiteratureReferencesDescriptiver.py
	11_rFamLiteratureReferencesTextMiningReporter.py
	12_rFam3DStructureDataDownloader.py
	13_rFam3DStructureDataEnrichment.py
	14_rFam3DStructureDataDescriptiver.py
<i>zinc_part</i> (ZINC veri tabanı kısmı)	0_dbDownloader.py
	1_dbBuilder.py
	2_ligandsBuilder.py
<i>docking_part</i> (Sana tarama analizi kısmı)	3_dbVisualiserAndDescriptiver.py
	0_vsReceptorSelectorAndPreparer.py
	1_vsAccuracyCalculator.py
	2_vsJobsRunner.py
	3_vsResultsReporter.py

Ayrıca geliştirilen ardışık düzene ait girdi ve çıktı dosyalarının anlaşılır bir biçimde arşivlenmesi için Tablo 5’te verilen özel bir “klasör adı kodlama” tekniği uygulanmıştır. Böylelikle kod betiklerinin yazım sürecinde gerekli girdi ve çıktı dosyalarının konumlarına ulaşılması ve takibi kolaylaştırılmıştır. Yine bu “klasör adı kodlama” tekniği bize her bir kod

betiğine ait tutulması gereken “log” kayıtlarının da düzenli bir şekilde tutulması konusunda yardımcı olmuştur.

Tablo 5. Kod betik dizinleri için kullanılan özel dosyalama tekniği içerikleri

		Klasörün	
Kodu	Adı	Kodu	Adı
3D	<i>3d_structure</i>	N	<i>ncbi</i>
3DP	<i>3d_structure_pool</i>	P	<i>plot</i>
A	<i>alignment</i>	PS	<i>property_space</i>
AN	<i>annotation</i>	RC	<i>rcsb</i>
B	<i>blast</i>	RCA	<i>rna_central_api</i>
D	<i>data</i>	RCND	<i>receptor_candidates</i>
DE	<i>descriptives</i>	RF	<i>rfam</i>
DP	<i>docking_part</i>	RP	<i>rfam_part</i>
E	<i>ensembl</i>	RSE	<i>receptor_selection</i>
F	<i>family</i>	RT	<i>riboswitch</i>
FA	<i>fa_files</i>	RWP	<i>raw_protomers</i>
FP	<i>fasta_pool</i>	SB	<i>sunburst</i>
FS	<i>ftp_server</i>	SML	<i>sample_ligands</i>
G	<i>gff</i>	SMR	<i>sample_results</i>
GR	<i>gc_ratio</i>	ST	<i>stats</i>
GT	<i>genome_type</i>	T	<i>taxonomy</i>
HR	<i>html_report</i>	TF	<i>tmp_files</i>
HS	<i>human</i>	TM	<i>text_mining</i>
ICC	<i>intra_class_cor</i>	VA	<i>vs_accuracy</i>
L	<i>log</i>	VR	<i>vs_results</i>
LB	<i>local_blast</i>	WR	<i>web_run</i>
LF	<i>lit_refs</i>	WST	<i>webina</i>
LP	<i>ligands_pool</i>	ZP	<i>zinc_part</i>
LR	<i>local_run</i>		

Bu ön bilgilendirme açıklamasından sonra tez yazımına her bir başlık altındaki ardışık düzen adımlarını gerçekleştirmek için yazılan kod betiklerinin açıklanması ile devam edilecektir.

3.3. Rfam Veri Tabanı Kısım İçin Oluşturulan Kod Betikleri

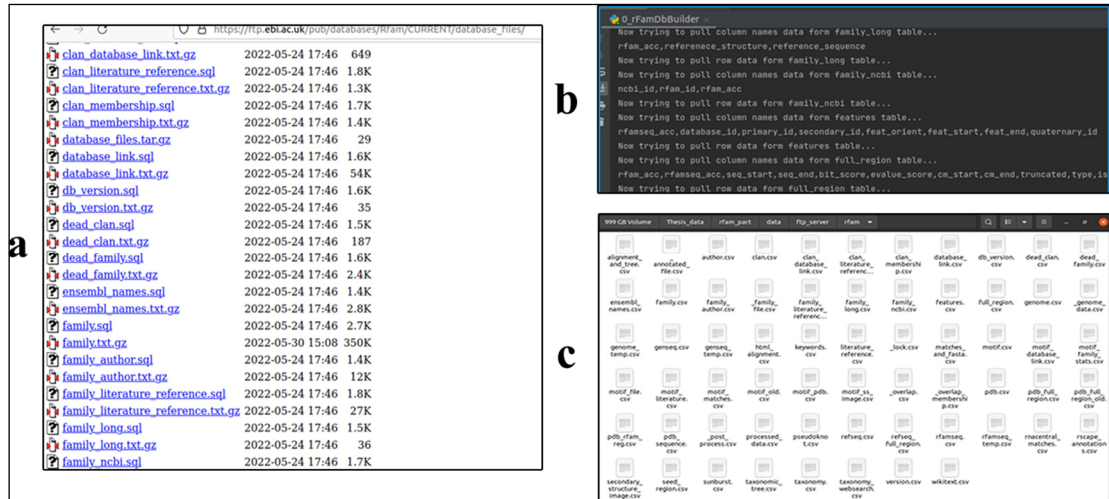
Bu kısımda oluşturulan kod betik dosya isimlerine ve kısaca ilgili dosyada ne yapıldığına ait açıklama için Tablo 6 incelenebilir. Detaylı açıklamalar aşağıda ilgili başlıklar altında verilmiştir.

Tablo 6. Rfam veri tabanı kısmı için oluşturulan kod betikleri

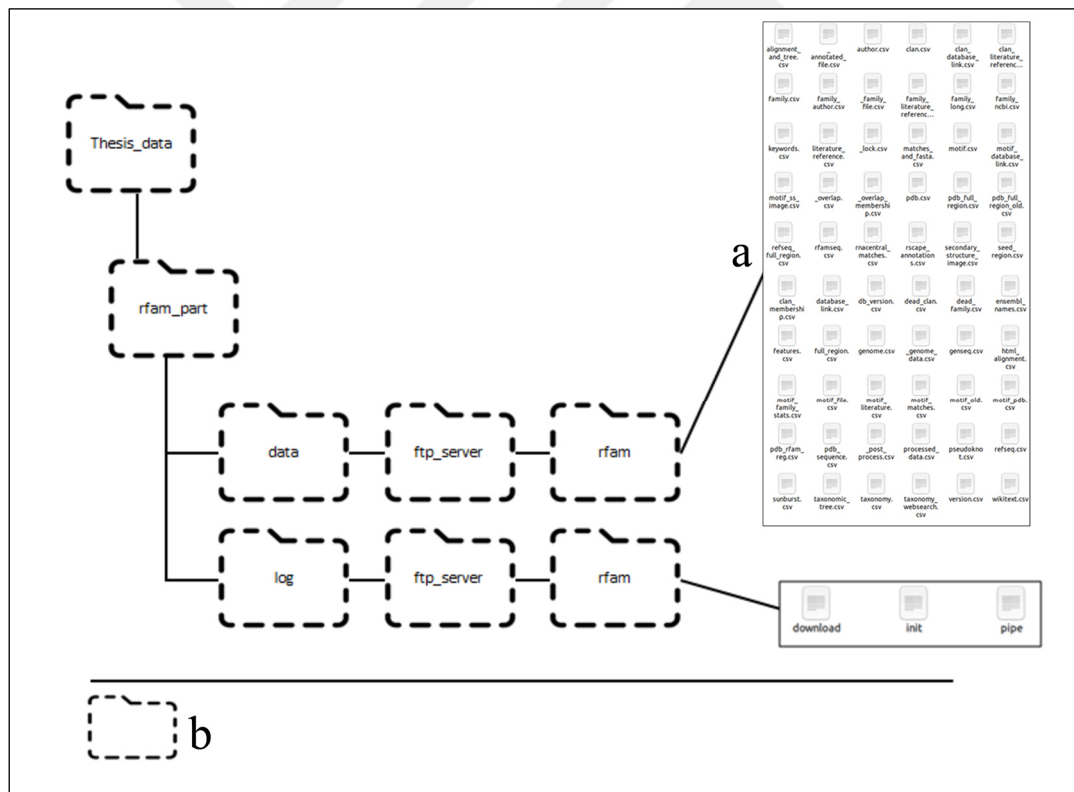
Kod betiğinin adı	Yaptığı işlem
0_rFamDbBuilder.py	Rfam SQL veri tabanı tablolarının yerel diske indirilmesi
1_rFamFamilyMembersDetector.py	Araştırma kapsamına giren Rfam aile türüne ait üye RNA elemanlarının belirlenmesi
2_rFamAlignmentDataBuilder.py	Belirlenen aile üyelerine ait hizalama verilerinin derlenmesi
3_rFamFASTADownloader.py	Belirlenen aile üyeleriyle ilgili FASTA dosyalarının indirilmesi
4_rFamGCRatioAnalyser.py	Belirlenen aile üyeleriyle ilgili indirilen FASTA dosyaları içerisinde GC oranı tespiti
5_rFamBlaster.py	İnsan genomu içerisinde belirlenen aile üyelerinin sekanslarının aranması
6_rFamBlastResultsAnnotator.py	BLAST sonuçlarının analiz edilmesi, eğer varsa ek açıklamaya değer kayıtların tespiti
7_rFamHumanGenomeAnnotationsDescriptiver.py	İnsan genomuna ait GFF dosyasının analiz edilmesi
8_rFamTaxonomicDataPuller.py	Belirlenen aile üyeleriyle ilgili taksonomik verilerin tespit edilmesi
9_rFamTaxonomicDataPlotter.py	Tespit edilen taksonomik verilerin grafik gösterimi
10_rFamLiteratureReferencesDescriptiver.py	Belirlenen aile üyeleriyle ilgili literatür referansı verilerinin tespit edilmesi
11_rFamLiteratureReferencesTextMiningReporter.py	Tespit edilen literatür referansı içeriklerine metin madenciliği uygulanması
12_rFam3DStructureDataDownloader.py	Belirlenen aile üyeleriyle ilgili üç boyutlu yapı dosyalarının yerel diske indirilmesi
13_rFam3DStructureDataEnrichment.py	İndirilen üç boyutlu yapı dosyalarına dair ek verilerin bulunması
14_rFam3DStructureDataDescriptiver.py	İndirilen üç boyutlu yapı dosyalarının bulunan ek verilere göre dağılımlarının tespiti

3.3.1. Rfam Veri Setinin Oluşturulması

Bu kısım için yazılan kod betiğinin yaptığı işlem erişime açık olan Rfam SQL veri tabanının birebir bir kopyasını belirtilen disk konumuna (yazara ait olan yerel bilgisayar ve KTÜ Tıp Bilişimi ABD sunucu bilgisayarı) yazılmasıdır. Kod betiği önce erişime açık olan Rfam SQL veri tabanı tablolarını tespit eder ve ardından her bir tablodaki verileri sırasıyla belirtilen disk konumuna satır satır yazar. Bu işlem yapılırken mevcut veri tabanının sürümüne de dikkat edilir. Eğer belirtilen disk konumunda zaten mevcut sürümüne ait tablolar varsa herhangi bir işlem yapılmaz. Ancak yeni bir sürüm olduğunda mevcut tabloları güncelleyerek eşitleme yapılır. İfade etmeye çalıştığımız süreç hakkında daha iyi fikir sahibi olabilmek için Resim 2 ve Şekil 2'yi incelemeniz önerilir.



Resim 2. *Rfam* veri tabanı oluşturma sürecinde kullanılan kod betiğinin yaptığı işlemler. **(a)** Erişime açık olan *Rfam* biyolojik SQL veri tabanı tabloları **(b)** Kod betiğinin çalışma sürecinden kesit **(c)** Kod betiği işlemini tamamladığında elde edilen çıktı dosyaları

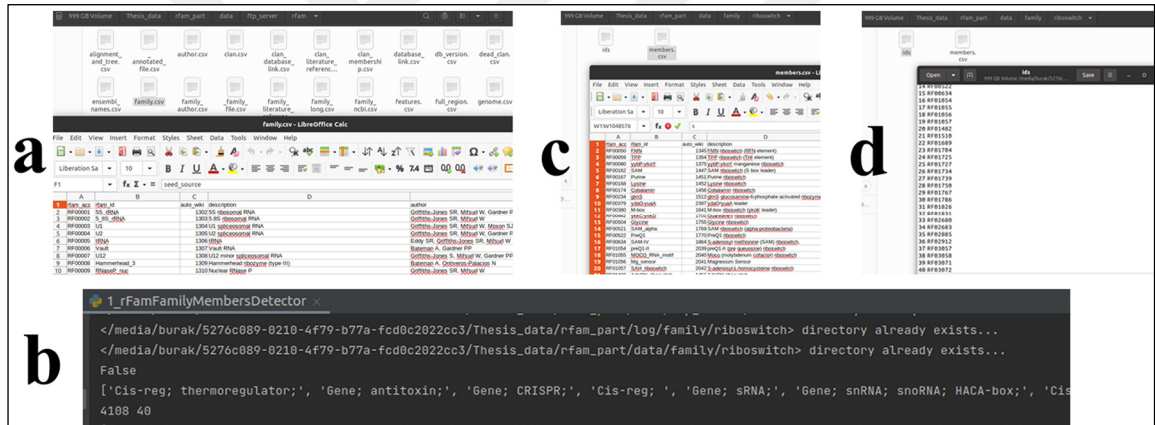


Şekil 2. *Rfam* veri tabanı oluşturma süreci için tasarlanan dizin yapısı ve elde edilen dosyalar. **(a)** Erişime açık olan *Rfam* biyolojik SQL veri tabanı tablolarının yerel bilgisayara aktarılmış kopyaları ve bu süreç için tutulan *log* kayıt dosyaları. **(b)** Kod betiğinin çalışma sürecinde belirtilen dizin için ilk defa oluşturulan klasörleri ifade eder.

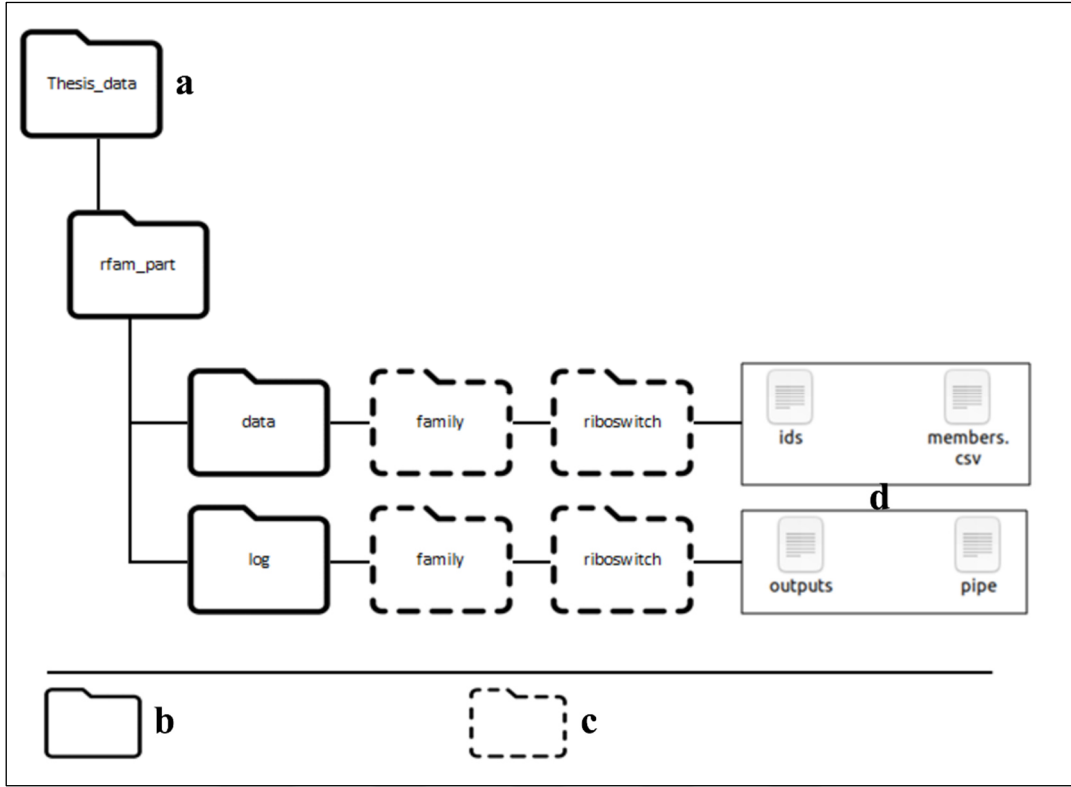
3.3.2. Riboşalter RNA Ailesine Ait Üyelerin Rfam Biyolojik Veri Tabanına Özel Kimlik Numaralarının Tespit Edilmesi

Bu kısım için yazılan kod betiğinin yaptığı işlem şu şekilde ifade edilebilir. Şekil 2’de gösterilen dizin yapısına aktarılmış olan Rfam biyolojik veritabanı dosyalarından “family.csv” dosyası üzerinde çalışma konumuz olan Riboşalter ile alakalı satırların filtrelenerek “members.csv” dosyası oluşturulması. Filtrelenmiş kayıtlara ait Rfam biyolojik veri tabanına ait özel kimlik numaralarının ise “ids” adlı dosyaya ayrıca kaydedilmesi.

Sonraki adımların tamamında uygulanacak işlemler “ids” olarak adlandırılan bu dosya içerisinde yer alan özel kimlik numarasına sahip Rfam üyeleri üzerinden yapılacaktır. İfade etmeye çalıştığımız süreç hakkında daha iyi fikir sahibi olabilmek için Resim 3 ve Şekil 3’ü incelemeniz önerilir.



Resim 3. Riboşalter RNA ailesine ait üyelerin Rfam biyolojik veri tabanına özel kimlik numaralarının tespit edilmesi sürecinde kullanılan kod betiğinin yaptığı işlemler. (a) Rfam biyolojik SQL veri tabanı tablolarından “family.csv” tablosunun içeriğinden bir kesit (b) Kod betiğinin çalışma sürecinden kesit (c) Kod betiği işlemini tamamladığında elde edilen çıktı dosyalarından “members.csv” dosya içeriğinden bir kesit (d) Kod betiği işlemini tamamladığında elde edilen çıktı dosyalarından “ids” dosya içeriğinden bir kesit



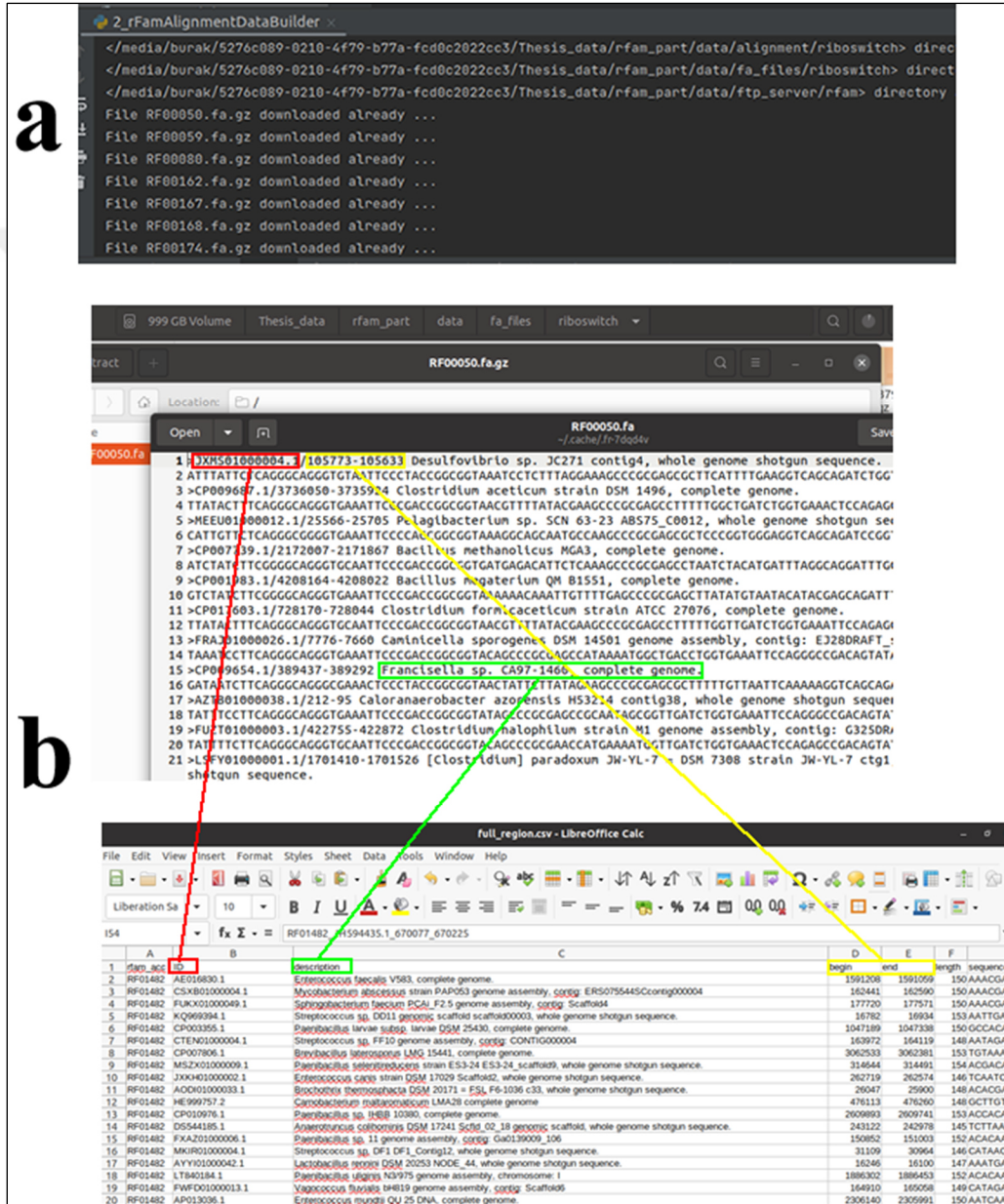
Şekil 3. Riboşalter RNA ailesine ait üyelerin *Rfam* biyolojik veri tabanına özel kimlik numaralarının tespit edilmesi süreci için tasarlanan dizin yapısı ve elde edilen dosyalar. **(a)** Kök dizini ifade eder. **(b)** Kod betiğinin çalışma sürecinde belirtilen kök dizin içerisinde daha önceki adımlarda oluşturulmuş klasörleri ifade eder. **(c)** Kod betiğinin çalışma sürecinde belirtilen kök dizin için ilk defa oluşturulan klasörleri ifade eder. **(d)** Kod betiği işlemini tamamladığında elde edilen çıktı ve *log* kayıt dosyalarını ifade eder.

3.3.3. Riboşalter RNA Ailesine Ait Üyeler İçin Hizalama Veri Setinin Oluşturulması

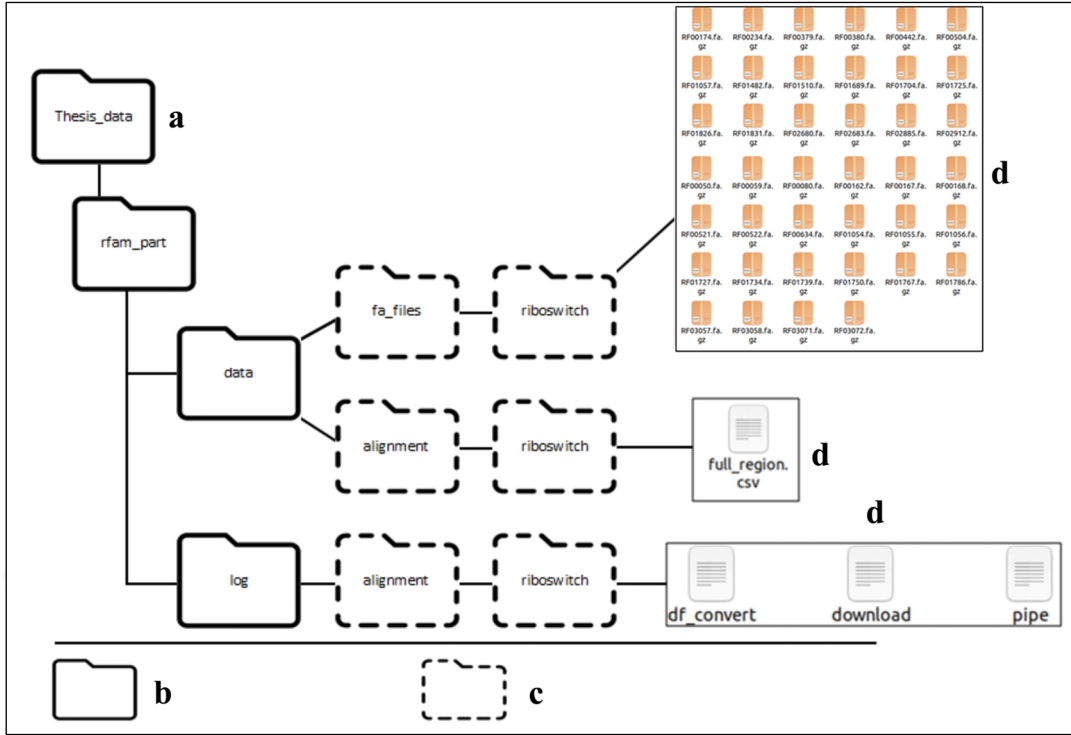
Erişime açık olan Rfam ftp sunucu içerisinde RNA elamanlarının tespit edildiği organizmalara ait sekans verileri de tutulmaktadır. FASTA formatının diğer bir versiyonu olan “*fa.gz*” formatında saklanan bu dosyalar her bir Rfam üyesine ait erişim numarası olarak adlandırılan dosyalar halinde sunucuda saklanmaktadır. Çalışmanın bu kısmında ilk olarak Riboşalter ailesi üyelerine ait erişim numarası ile ilgili olan 40 adet “*fa.gz*” dosyaları sunucudan lokal diske indirilmiştir.

Bahsi geçen dosyaların içeriğinde çoklu Fasta kayıtlarını içeren “seed” hizalama sekans verileri bulunmaktadır. Ardından lokal diskte yer alan bahsi geçen bu dosyalar “byte” düzeyinde okunarak (herhangi dışarı çıkarma işlemine gerek olmadan), içerisinden çalışma için gerekli olan veriler ilgili “Phyton” paketini kullanarak belirlediğimiz sütunlara (*ID*, *description*, *begin*, *end*, *length*, *sequence*) metin işleme yöntemiyle aktarılmıştır (92). Bu

sütunlara ilave olarak çalışmamız için oluşturduğumuz özel veri tabanı kayıtları için “rfam_acc” sütunu da bunlara ilave edilmiştir. Belirtilen sütunlara sahip olan dosya “full_region.csv” adı ile lokal diske kaydedilmiş ve tez çalışmasının sonraki adımlarında da girdi dosyası olarak kullanılmıştır. İfade etmeye çalıştığımız süreç hakkında daha iyi fikir sahibi olabilmek için Resim 4 ve Şekil 4’ü incelemeniz önerilir.



Resim 4. Riboşalter RNA ailesine ait üyeler için hizalama veri setinin oluşturulması sürecinde kullanılan kod betiğinin yaptığı işlemler. (a) Kod betiğinin çalışma sürecinden kesit (b) Kod betiği işlemini tamamladığında elde edilen “full_region.csv” dosya içeriğinin oluşturulma sürecini betimleyen bir kesit.

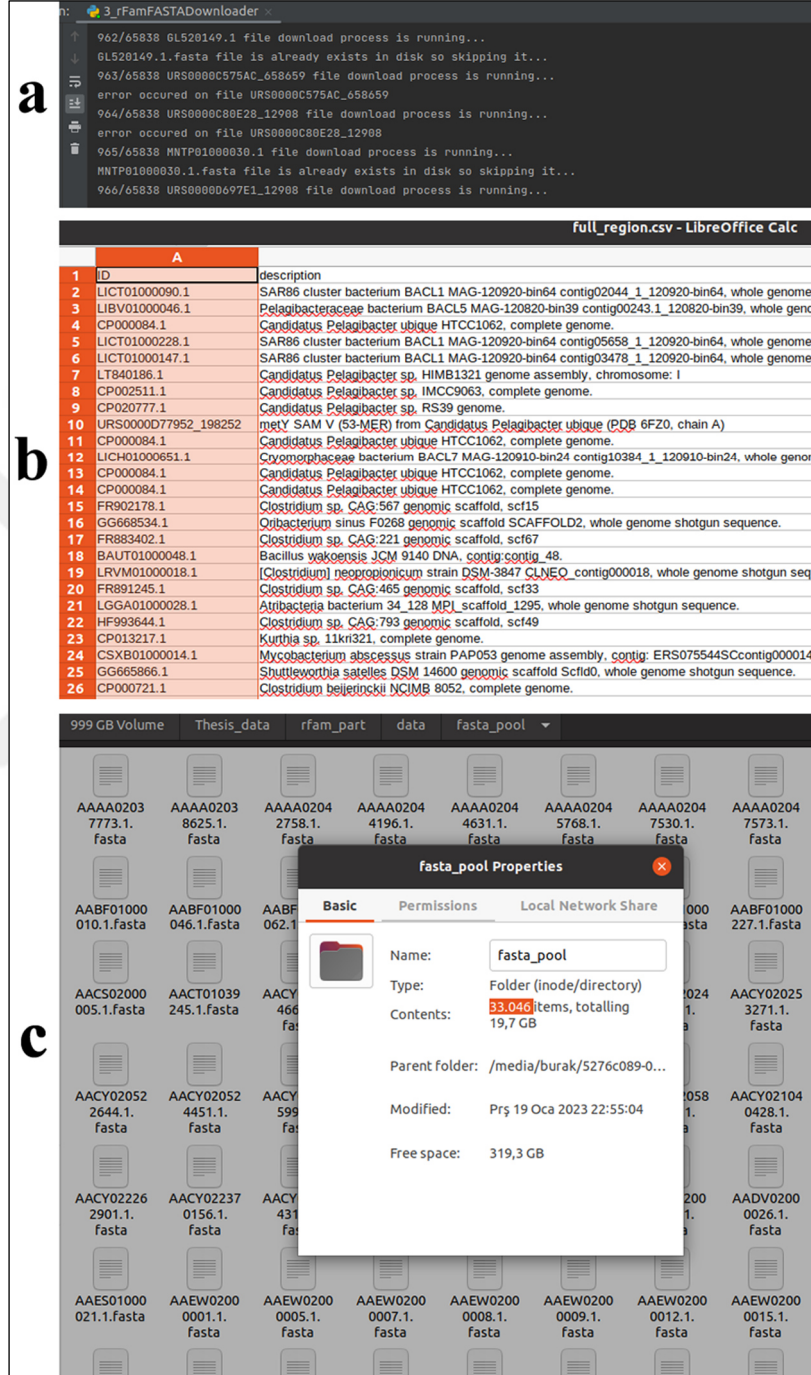


Şekil 4. Ribosalter RNA ailesine ait üyeler için hizalama veri setinin oluşturulması süreci için tasarlanan dizin yapısı ve elde edilen dosyalar. **(a)** Kök dizini ifade eder. **(b)** Kod betiğinin çalışma sürecinde belirtilen kök dizin içerisinde daha önceki adımlarda oluşturulmuş klasörleri ifade eder. **(c)** Kod betiğinin çalışma sürecinde belirtilen kök dizin için ilk defa oluşturulan klasörleri ifade eder. **(d)** Kod betiği işlemini tamamladığında elde edilen çıktı ve *log* kayıt dosyalarını ifade eder.

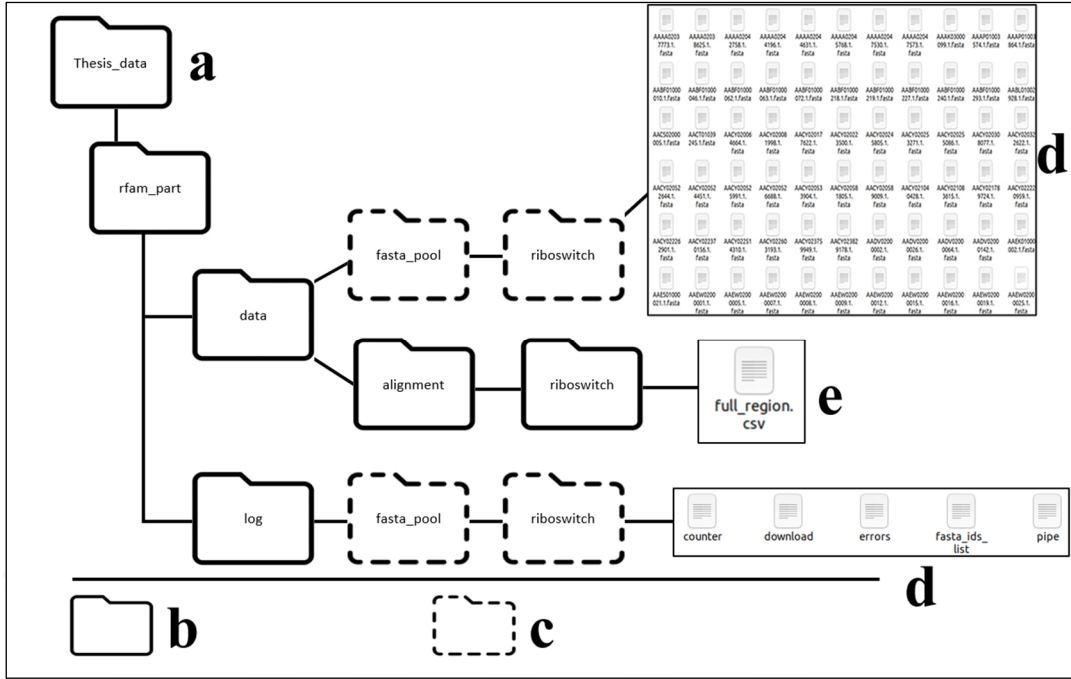
3.3.4. Ribosalter RNA Ailesine Ait Üyeler İçin FASTA Dosyalarının İndirilmesi

Bir önceki adımda elde edilmiş olan “*full_region.csv*” içerisinde hizalanmış sekans verileri bulunmaktadır. Bu kısımda yapılan işlem ise hizalanmış sekans verilerinin ait olduğu kaynak “*Fasta*” dosyalarının yerel diske indirilmesidir. Çalışmanın sonraki analizlerinin birçok kısmında indirilen bu “*Fasta*” dosyaları kullanılmıştır. “*BioPhyton*”(92) paketine ait kütüphaneler yardımıyla “*full_region.csv*” dosyasında yer alan “*Fasta*” dosyalarına ait NCBI biyolojik veri tabanına ait özel erişim kimlik numaraları kullanılarak 33 046 adet dosya diske indirilmiştir. Bu kısımdaki işlemler için girdi verisi sunan “*full_region.csv*” dosyası incelendiğinde içerisinde 65 838 adet kaydın olduğu görülmektedir. İndirilen dosya sayısının burada ifade edilen kayıt sayısından farklı olmasının nedeni, kayıtlardan bazılarının aynı genomu ait “*Fasta*” dosyasının farklı kısımları ile elde edilmiş hizalama sekanslarına ait olmasıdır. Ayrıca, 3 732 adet dosyanın ise özel erişim kimlik numarası artık kullanılmadığı için indirilememiştir. İndirilemeyen bu dosyalar için “*log*” kayıtları ayrıca

tutulmuştur. İfade etmeye çalıştığımız süreç hakkında daha iyi fikir sahibi olabilmek için Resim 5 ve Şekil 5’i incelemeniz önerilir.



Resim 5. Riboşalter RNA ailesine ait üyeler için FASTA dosyalarının indirilmesi sürecinde kullanılan kod betiğinin yaptığı işlemler. (a) Kod betiğinin çalışma sürecinden kesit (b) Kod betiği işleminde girdi verisi olarak kullanılan “full_region.csv” dosya içeriğinin ilgili sütununu gösteren bir kesit. (c) Kod betiği çalışmasını tamamladığında yerel diske indirilen dosyaları gösteren bir kesit.



Şekil 5. Ribosalter RNA ailesine ait üyeler için FASTA dosyalarının indirilmesi süreci için tasarlanan dizin yapısı ve elde edilen dosyalar. **(a)** Kök dizini ifade eder. **(b)** Kod betiğinin çalışma sürecinde belirtilen kök dizin içerisinde daha önceki adımlarda oluşturulmuş klasörleri ifade eder. **(c)** Kod betiğinin çalışma sürecinde belirtilen kök dizin için ilk defa oluşturulan klasörleri ifade eder. **(d)** Kod betiği işlemini tamamladığında elde edilen çıktı ve *log* kayıt dosyalarını ifade eder. **(e)** Kod betiğinin çalışması için gerekli girdi dosyasını ifade eder.

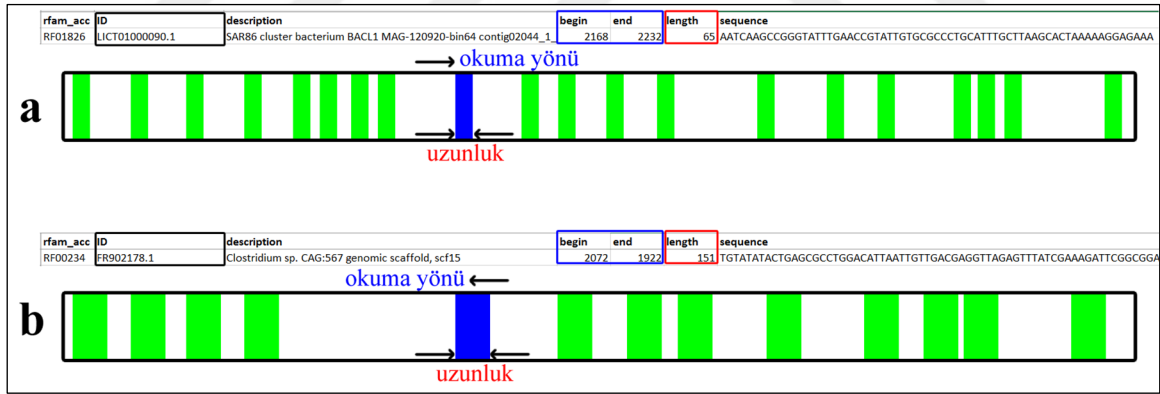
3.3.5. Ribosalter RNA Ailesine Ait Üyeler İçin GC-Oranı Analizlerinin Yapılması

Çalışmanın bu kısmında “*full_region.csv*” dosyası içerisinde bulunan hizalama sekans verileri kullanılarak temel bir biyoinformatik analizi olan GC-Oranı’nın yapılabilmesi için önce gerekli veri setleri oluşturulmuştur. Her bir ribosalter aile üyesi için oluşturulan veri setlerinde süreç aşağıda belirtilen akış üzerinden işletilmiştir:

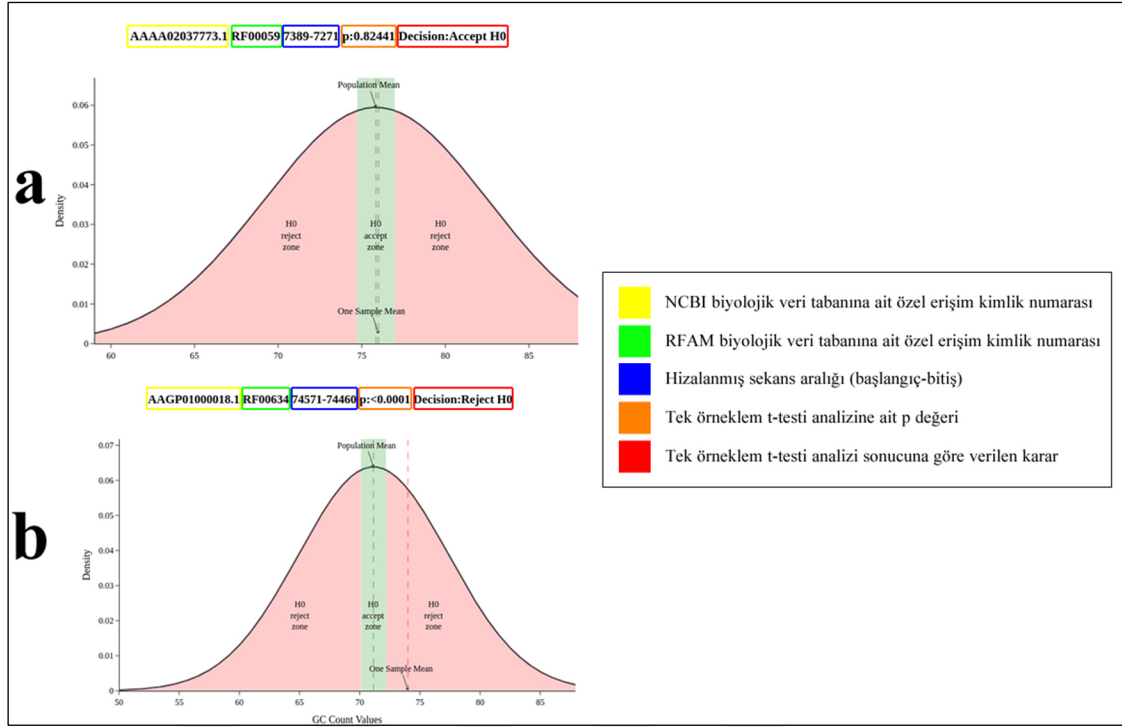
- 1) Girdi verisi olan “*full_region.csv*” dosyası içerisindeki kayıt satırlarında yer alan hizalama sekans verisi üzerinden *BioPhyton* (92) paketinin ilgili kütüphanesinde yer alan fonksiyon yardımıyla her bir sekansa ait GC-oranının hesaplanması (tek örneklem) ve “*results.csv*” dosyasına kaydedilmesi.
- 2) Ardından Resim 6’da gösterilen yöntem ile “*results.csv*” dosyasında bulunan her bir sekansın yer aldığı genoma ait FASTA dosyası kullanılarak hizalama sekansı ile bire bir aynı özellikte (sekansın uzunluğu ve okunma şekli) rastgele seçilen 100 farklı hizalama sekans örnekleme seçilmiştir ve bu örnekleme ait GC-oranları hesaplanmıştır (popülasyon).

- 3) Ardından elde edilen örneklemin popülasyon kabul edildiği ve “*results.csv*” dosyasındaki kayıtlarda yer alan hizalama sekansını ise tek örneklem değeri kabul eden tek örneklem t istatistiği hesaplanmıştır. Bu hesaplama her bir kayıt için benzer mantıkla elde edilen istatistik sonuçları Resim 7’de gösterilen şekilde grafik çıktısı olarak düzenlenmiştir.
- 4) Tek örneklem t istatistiğine ait analiz sonuçları ayrıca “*reports.csv*” dosyasına da yazılmış ve bu dosya grafik çıktıları da barındıracak şekilde Resim 8’de gösterildiği gibi “*html*” çıktısı olarak raporlanmıştır.
- 5) Son adımda ise her bir ribozalter aile üyesine ait istatistik sonuçlarını toplu halde yorumlayabilmek için tüm istatistik analizleri sonucunda elde edilen *p* değerlerinin dağılımları hesaplanmıştır. Elde edilen çıktı dosyası “*descriptive.csv*” adıyla ilgili dizine yazılmıştır. Bu dosyanın tablo düzenine uyarlanmış hali tezin bulgular kısmında verilmiş ve yorumlanmıştır.

Burada ifade edilen işlem adımlarına ait süreçlerin daha iyi anlaşılması için Resim 9 ve Şekil 6’yı incelemeniz önerilir.



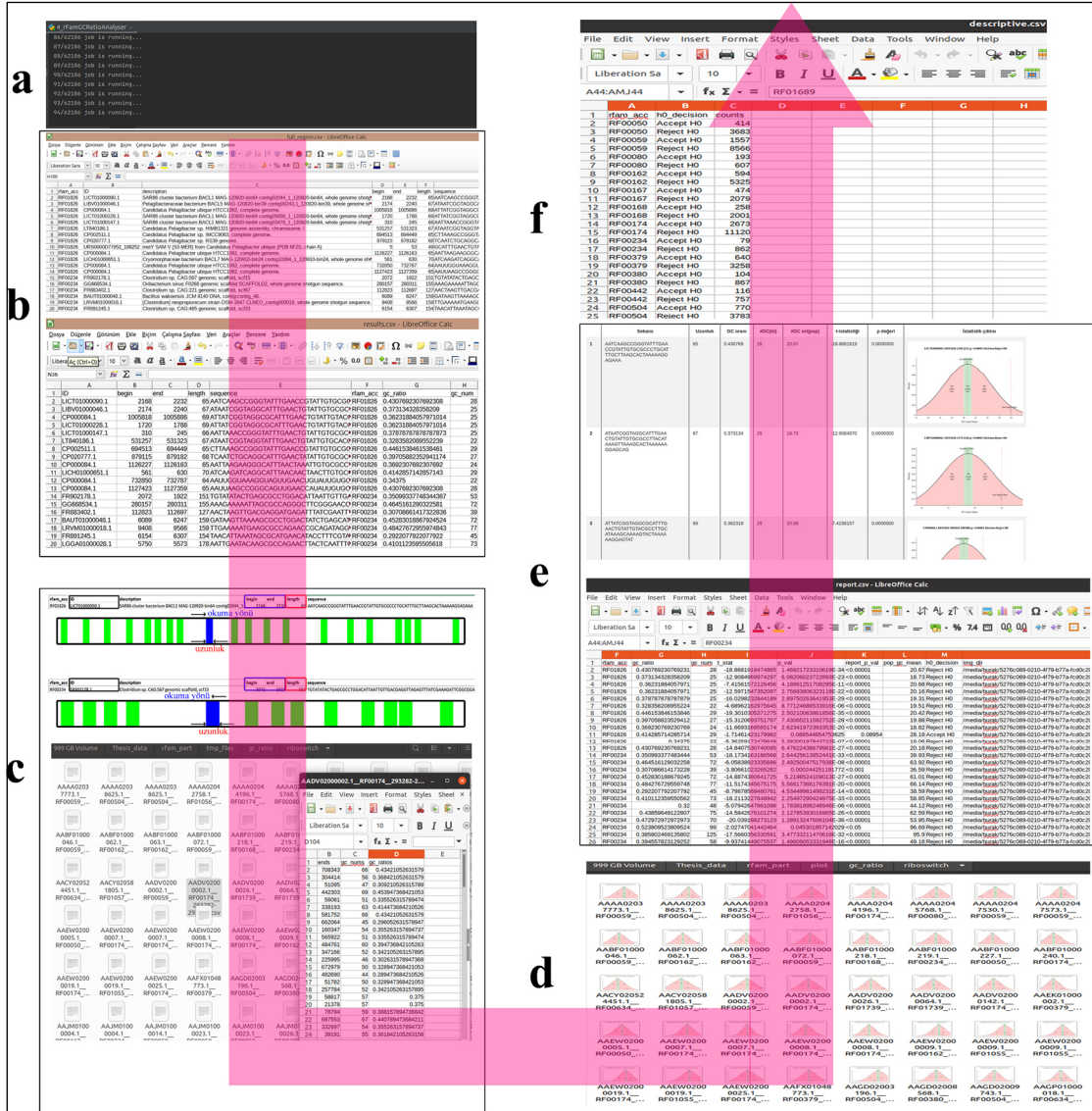
Resim 6. GC-Oranı analizlerine ait ifade edilen işlem akışındaki ikinci adımda ifade edilen sekans uzunluğu ve sekansın okunma şekli ifadelerinin betimlenmesi. **(a)** Normal sekans okunma sürecini ifade eder. **(b)** Tersten sekans okunma sürecini ifade eder. Mavi renkli dikdörtgen tek örneklem sekansını yeşil renkliler ise rastgele seçilenleri yani popülasyon sekanslarını ifade eder. Siyah renkli büyük dikdörtgenler ise okunan FASTA dosyalarını ifade eder.



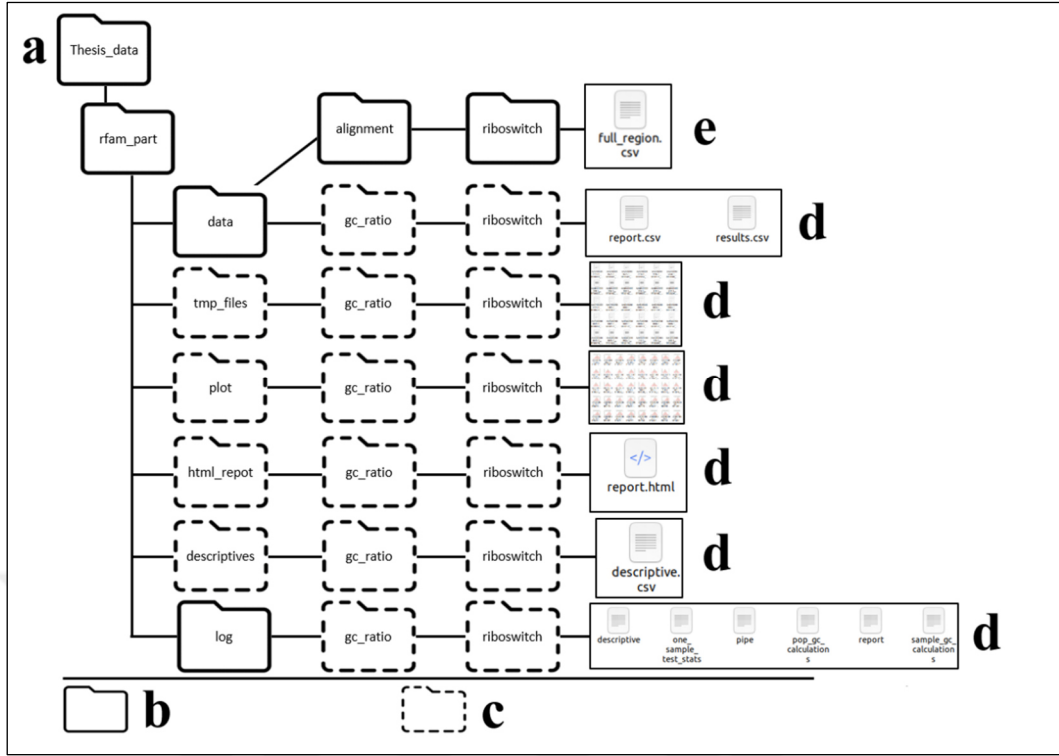
Resim 7. GC-Oranı analizlerine ait ifade edilen işlem akışındaki üçüncü adımda ifade edilen tek örneklem t istatistiği hesaplanma sürecinin betimlenmesi. **(a)** H_0 hipotezinin kabul edildiği durumu ifade eder. **(b)** H_0 hipotezinin reddedildiği durumu ifade eder.

	Sekans	Uzunluk	GC oranı	#GC(tö)	#GC ort(pop)	t-istatistiği	p değeri	İstatistik çıktısı
1	AATCAAGCCGGTATTGAA CCGTATTGCGCCCTGCAT TTGCTTAAGCACTAAAAGG AGAAA	65	0.430769	28	20.67	-18.8681919	0.0000000	
62105	CGUAUCAUUGGCCUCGCCU GCGCCGGAAGUCCUAGGAC AUCCCGCGCCCGCGGGGCG AUAAGUCGAAAGAGGGAUG AUAUGAC	87	0.620690	54	59.21	9.2413462	0.0000000	

Resim 8. GC-Oranı analizlerine ait ifade edilen işlem akışındaki dördüncü adımda ifade edilen “html” rapor çıktısından bir kesit.



Resim 9. Riboşalter RNA ailesine ait üyeler için GC-oranı analizi sürecinde kullanılan kod betiğinin yaptığı işlemler. **(a)** Kod betiğinin çalışması sürecinden kesit **(b)** Kod betiği işleminde girdi verisi olarak kullanılan “full_region.csv” dosya içeriği kullanılarak elde edilen “results.csv” dosyasını gösteren bir kesit. **(c)** “results.csv” dosyasında tek örneklem için hesaplanan GC oranı değerinin 100 örneklem ile temsil edilen popülasyonlar içinde hesaplanma sürecini gösteren kesit. **(d)** Tek örneklem t istatistiği sonuçlarını grafik çıktı olarak elde edilmesi sürecinden kesit. **(e)** GC-oranı analizlerine ait p-değerlerinin her bir riboşalter aile üyesi üzerine dağılımını veren “descriptives.csv” dosyasından bir kesit.



Şekil 6. Ribosalter RNA ailesine ait üyeler için GC-oranı analizi süreci için tasarlanan dizin yapısı ve elde edilen dosyalar. **(a)** Kök dizini ifade eder. **(b)** Kod betiğinin çalışma sürecinde belirtilen kök dizin içerisinde daha önceki adımlarda oluşturulmuş klasörleri ifade eder. **(c)** Kod betiğinin çalışma sürecinde belirtilen kök dizin için ilk defa oluşturulan klasörleri ifade eder. **(d)** Kod betiği işlemini tamamladığında elde edilen çıktı, grafik, rapor ve *log* kayıt dosyalarını ifade eder. **(e)** Kod betiğinin çalışması için gerekli girdi dosyasını ifade eder.

3.3.6. Ribosalter RNA Ailesine Ait Üyelerin Hizalanmış Sekanslarının İnsan Genomunda Aranması

Bu kısımda yapılan çalışmalar iki alt başlıkta verilmiştir. Birinci başlıkta ribosalter RNA aile üyelerine ait hizalanmış sekansların NCBI biyolojik veri tabanında yer alan BLAST analizi ile ön analizinin yapılma süreci anlatılmıştır.

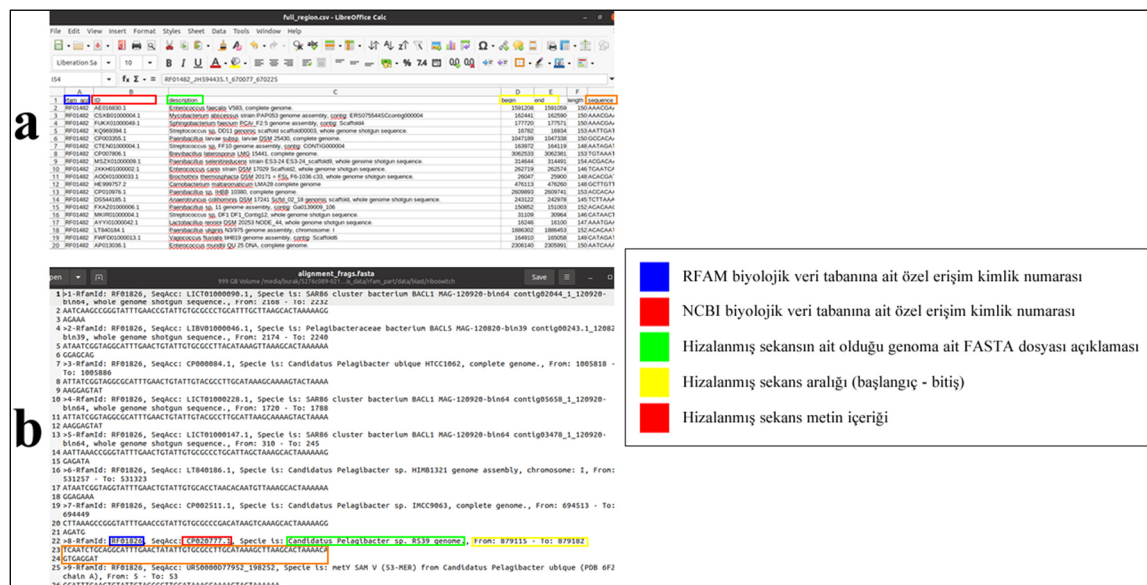
İkinci alt başlıkta ise anlamlı eşleşme bulunan sekansların literatürde *anotasyon* bilgisi olup olmadığının RNA Central biyolojik veri tabanında yer alan *web servisi (api)* yardımıyla araştırılması sürecinde yapılan işlemler anlatılmıştır.

3.3.6.1. NCBI Biyolojik Veri Tabanı Üzerinden BLAST Analizi Süreci

Çalışmanın bu kısmında “*full_region.csv*” dosyası içerisinde bulunan hizalama sekans verileri kullanılarak ribosalter aile üyelerine ait çeşitli genom sekansları içerisinde anlamlı eşleşme bulunarak raporlanmış hizalama sekanslarının toplu bir FASTA dosyasına

dönüştürülmesi ve bu sekansların NCBI biyolojik veri tabanında insan (*Homo sapiens*) genomu içerisinde BLAST analizi ile aranması süreci ifade edilmiştir. Bu kısım için oluşturulan kod betiğinin gerçekleştirdiği işlem adımlarına ait akış süreci aşağıda ifade edilmiştir:

Birinci adım: Girdi verisi olan “*full_region.csv*” dosyası içerisindeki kayıt satırlarında yer alan veriler, *BioPhyton* (92) paketinin ilgili kütüphanesinde yer alan fonksiyonlar yardımıyla Resim 10’da gösterilen düzende derlenerek “*alignment_fragments.fasta*” adlı dosyaya yazılmıştır.



Resim 10. BLAST analizi sürecinde girdi olarak kullanılan “*alignment_frgs.fasta*” dosyasının oluşturulma süreci. **(a)** Kod betiğinin düzgün çalışması için gerekli girdi dosyası olan “*full_region.csv*” içeriğinden bir kesit. **(b)** BLAST analizi için girdi olarak kullanılacak “*alignment_frgs.fasta*” dosya içeriğinden bir kesit.

İkinci adım: Oluşturulan “*alignment_fragments.fasta*” adlı dosyanın direkt olarak NCBI biyolojik veri tabanı üzerinden “*online*” BLAST aracına gönderilmesinin mümkün olmadığı (sunucunun uzun süre cevap veremediği ve hata döndürdüğü) gözlenmiştir. Bu nedenle bahsi geçen dosya “*PySplit*” paketinin ilgili kütüphanesinde yer alan fonksiyonlar yardımıyla daha küçük boyutlu dosyalara ayrılmıştır.

Üçüncü adım: Küçük parçalara ayrılmış FASTA dosyaları bir döngü içerisinde *BioPhyton* (92) paketinin ilgili kütüphanesinde yer alan fonksiyonlar yardımıyla “*online*” BLAST aracına analiz için gönderilmiştir. Her bir analiz sonucunda gelen “*xml*”

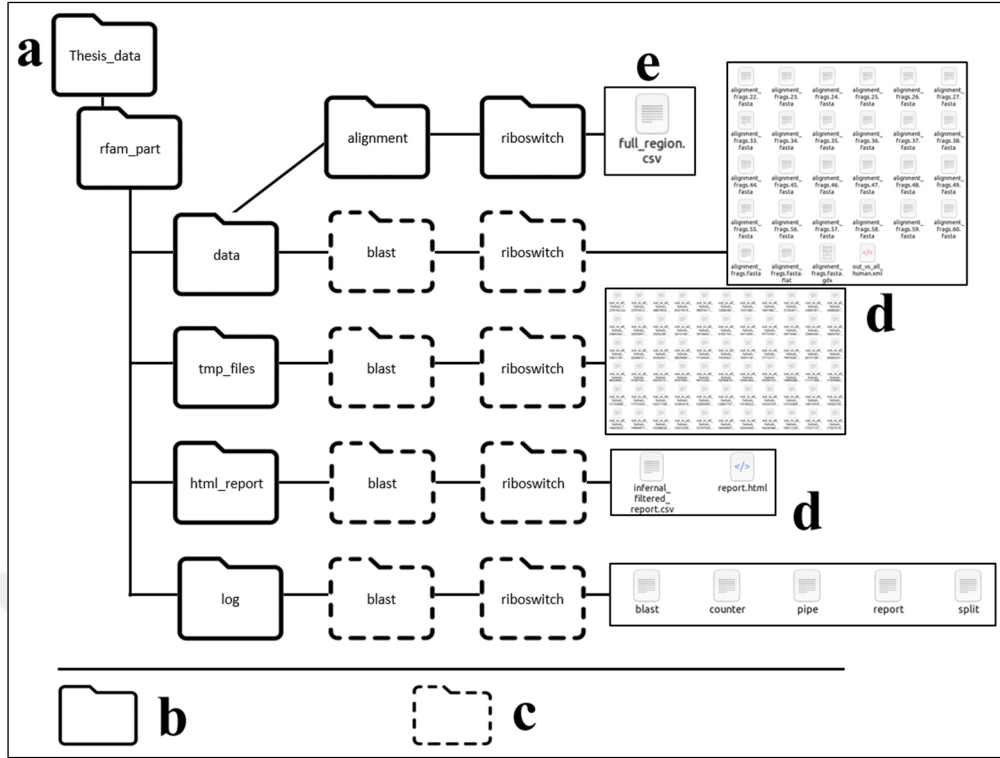
formatındaki sonuçlar aynı “xml” dosyası (*out_vs_all_human.xml*) üzerine yazılarak biriktirilmiştir.

	Query_definition	Hit_id	Hit_def	Hit_accession	Hit_len	Hsp_bit_score	Hsp_evalue	Hsp_query_from	Hsp_query_to	Hsp_hit_from	Hsp_hit_to
1	33319-Rfam: RF00168, SeqAcc: CP000423.1, Specie is: Lactobacillus casei ATCC 334, complete genome., From: 92631 - To: 92453	gi 272989689 gb GU266878.1	Homo sapiens contig freeze2_11051 genomic sequence	GU266878	1117	308.7610	2.892820e-81	1	179	32	211
2	34163-Rfam: RF00168, SeqAcc: AEYMI001934.1, Specie is: Lactobacillus rhamnosus MTCC 5462 contig45, whole genome shotgun sequence., From: 1053 - To: 874	gi 272989689 gb GU266878.1	Homo sapiens contig freeze2_11051 genomic sequence	GU266878	1117	222.1990	3.326090e-55	1	180	32	211
3	8795-Rfam: RF00379, SeqAcc: CP000024.1, Specie is: Thermotoga bacteriophages pseudethanolus ATCC 33223, complete genome., From: 1503502 - To: 1503372	gi 1781382119 pdb 6NST A	Chain A, RNA (126-MER)	6NST_A	126	205.9690	1.710480e-50	10	129	3	126
4	6474-Rfam: RF00379, SeqAcc: CP000024.1, Specie is: Thermotoga bacteriophages pseudethanolus ATCC 33223, complete genome., From: 1503502 - To: 1503372	gi 1781382119 pdb 6NST A	Chain A, RNA (126-MER)	6NST_A	126	205.9690	1.710500e-50	10	129	3	126
5	8795-Rfam: RF00379, SeqAcc: CP000024.1, Specie is: Thermotoga bacteriophages pseudethanolus ATCC 33223, complete genome., From: 1503502 - To: 1503372	gi 1781382113 pdb 6NSN A	Chain A, RNA (125-MER)	6NSN_A	125	198.7550	2.538580e-48	10	129	3	125
6	8795-Rfam: RF00379, SeqAcc: CP000024.1, Specie is: Thermotoga bacteriophages pseudethanolus ATCC 33223, complete genome., From: 1503502 - To: 1503372	gi 1781382116 pdb 6NSQ A	Chain A, RNA (128-MER)	6NSQ_A	128	197.8540	2.538580e-48	10	129	3	128
183	8714-Rfam: RF00379, SeqAcc: FQZV01000016.1, Specie is: Geopora bacteriophages subterraneus DSM 17957 genome assembly, contig E.J56DRAFT_scaffolds0014.14, From: 37934 - To: 37799	gi 1781382115 pdb 6NSP A	Chain A, RNA (127-MER)	6NSP_A	127	73.4211	7.654040e-11	33	134	23	127
184	8714-Rfam: RF00379, SeqAcc: FQZV01000016.1, Specie is: Geopora bacteriophages subterraneus DSM 17957 genome assembly, contig E.J56DRAFT_scaffolds0014.14, From: 37934 - To: 37799	gi 1781382113 pdb 6NSN A	Chain A, RNA (125-MER)	6NSN_A	125	73.4211	7.654040e-11	33	134	21	125
185	8714-Rfam: RF00379, SeqAcc: FQZV01000016.1, Specie is: Geopora bacteriophages subterraneus DSM 17957 genome assembly, contig E.J56DRAFT_scaffolds0014.14, From: 37934 - To: 37799	gi 1781382112 pdb 6NSL A	Chain A, RNA (124-MER)	6NSL_A	124	73.4211	7.654040e-11	33	134	20	124
186	8714-Rfam: RF00379, SeqAcc: FQZV01000016.1, Specie is: Geopora bacteriophages subterraneus DSM 17957 genome assembly, contig E.J56DRAFT_scaffolds0014.14, From: 37934 - To: 37799	gi 2033712817 pdb 6WTR A	Chain A, RNA (125-MER)	6WTR_A	125	74.3227	7.654040e-11	10	134	3	125
187	8714-Rfam: RF00379, SeqAcc: FQZV01000016.1, Specie is: Geopora bacteriophages subterraneus DSM 17957 genome assembly, contig E.J56DRAFT_scaffolds0014.14, From: 37934 - To: 37799	gi 1781382114 pdb 6NSO A	Chain A, RNA (126-MER)	6NSO_A	126	73.4211	7.654040e-11	33	134	22	126

Resim 11. BLAST analizlerine ait ifade edilen işlem akışındaki dördüncü adımda ifade edilen “html” rapor çıktısından bir kesit.

Dördüncü adım: Tüm sonuçların yer aldığı “xml” dosyası çeşitli metin işleme fonksiyonları yardımıyla işlenerek, her bir döngü adımına ait kısımlar “csv” dosya formatına dönüştürülerek “tmp_files” klasörü altında saklanmıştır. Ardından bahsi geçen tüm “csv” dosyaları tek bir “csv” dosyası altında (*infernal_filtered_report.csv*) toplanmış ve Resim 10’da gösterildiği gibi “html” formatında bir rapor dosyası (*report.html*) oluşturularak bu kısma ait süreç tamamlanmıştır. Rapor dosyası içerisinde belirlenen eşik E-değerinin ($1e^{-10}$) altındaki kayıtlar bulunmaktadır.

Burada ifade edilen işlem adımlarına ait süreçlerin daha iyi anlaşılması için Resim 12 ve Şekil 7’yi incelemeniz önerilir.



Şekil 7. Ribosalter RNA ailesine ait üyeler için BLAST analizi süreci için tasarlanan dizin yapısı ve elde edilen dosyalar. **(a)** Kök dizini ifade eder. **(b)** Kod betiğinin çalışma sürecinde belirtilen kök dizin içerisinde daha önceki adımlarda oluşturulmuş klasörleri ifade eder. **(c)** Kod betiğinin çalışma sürecinde belirtilen kök dizin için ilk defa oluşturulan klasörleri ifade eder. **(d)** Kod betiği işlemini tamamladığında elde edilen çıktı, rapor ve *log* kayıt dosyalarını ifade eder. **(e)** Kod betiğinin çalışması için gerekli girdi dosyasını ifade eder.

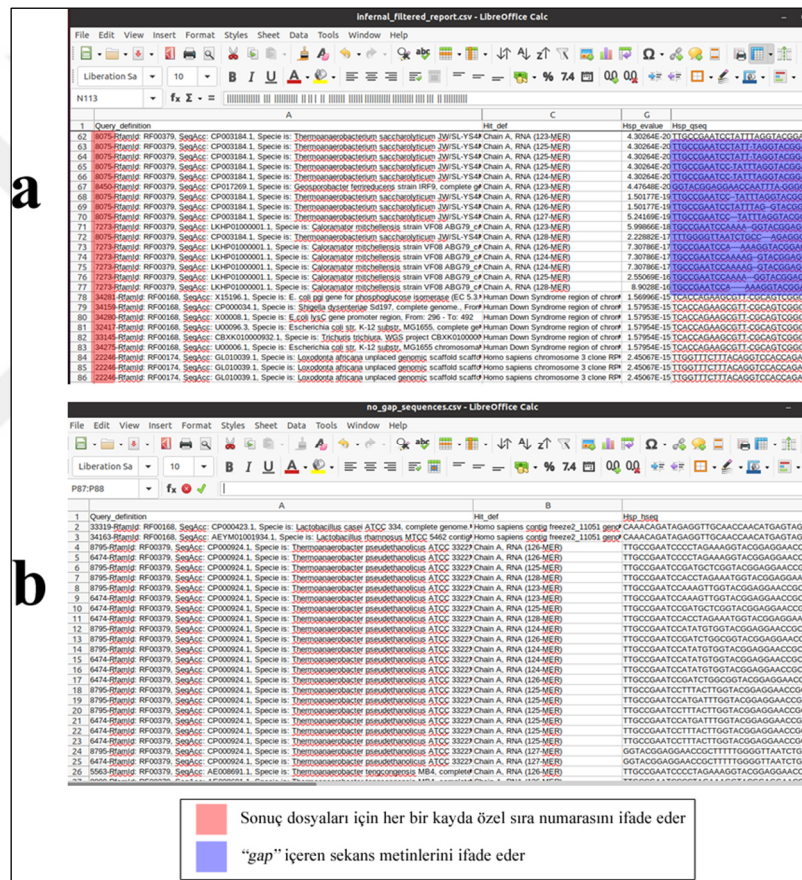
3.3.6.2. RNA Central Biyolojik Veri Tabanı Üzerinden “Anotasyon” Analizi Süreci

Çalışmanın bu kısmında “*infernai_filtered_report.csv*” dosyası içerisinde bulunan insan (*H. sapiens*) genomu ile anlamlı BLAST analizi eşleşmesine sahip sekans verilerine ait literatürde tanımlanmış anotasyon verisi olup olmadığı araştırılmıştır. Bu tez çalışmasının amaçlarından birisi de ribosalter adlı RNA elemanlarına benzer yapıların insan (*H. sapiens*) genomu içerisinde de bulunup bulunmadığının araştırılmasıydı. Çalışmanın bu kısmında izlenen yöntem ile tezin bu amacına yönelik bulguların da elde edilmesi hedeflenmiştir.

“*RNA Central*” veri tabanına ait web servisleri (*api*) kullanılarak anlamlı eşleşme bulduğumuz sekanslar, hem “*Rfam*” veri tabanında hem de çeşitli anotasyon veri kayıtları tutan biyolojik veri tabanlarında aranmıştır. Bulunan anlamlı anotasyon bilgileri hem “*Rfam*” biyolojik veri tabanına ait bulguları hem de diğer biyolojik veri tabanlarına ait

bulguları da kapsayacak şekilde detaylı olarak raporlanmıştır. Bu kısım için oluşturulan kod betiğinin gerçekleştirdiği işlem adımlarına ait akış süreci aşağıda ifade edilmiştir:

Birinci adım: Girdi verisi olan “*infernal_filtered_report.csv*” dosyası yer alan “*gap*” içeren sekans eşleşmeleri analiz kapsamı dışında bırakılması gerektiğinden bu türdeki kayıtlar dosyadan filtrelenerek “*no_gap_sequences.csv*” adlı dosya oluşturulur. Filtreleme işlemine gerek duyulmasının sebebi çalışmada kullandığımız RNA Central biyolojik veri tabanına ait web servisinin “*gap*” içeren sekanslar üzerinden işlem yapamıyor oluşudur. Bu adımda ifade edilen sürecin ve terimlerin daha iyi anlaşılması için Resim 13 incelenebilir.



Resim 13. Ribozalter RNA ailesine ait üyeler için anotasyon analizi sürecinde kullanılan kod betiğine ait akış sürecinin birinci adımında yapılan işlemler. (a) Girdi dosyası olan “*infernal_filtered_report.csv*” içeriğinden bir kesit. (b) Filtreleme işlemi sonucunda elde edilen “*no_gap_sequences.csv*” adlı dosya içeriğinden bir kesit.

İkinci adım: Bu adımda filtreleme yöntemi ile elde edilen “*no_gap_sequences.csv*” dosyasında yer alan her bir kayıt için sekans arama analizleri yapılmıştır. Bu süreç için <https://rnacentral.org/sequence-search/api> adresinde bulunan insan etkileşimine açık web

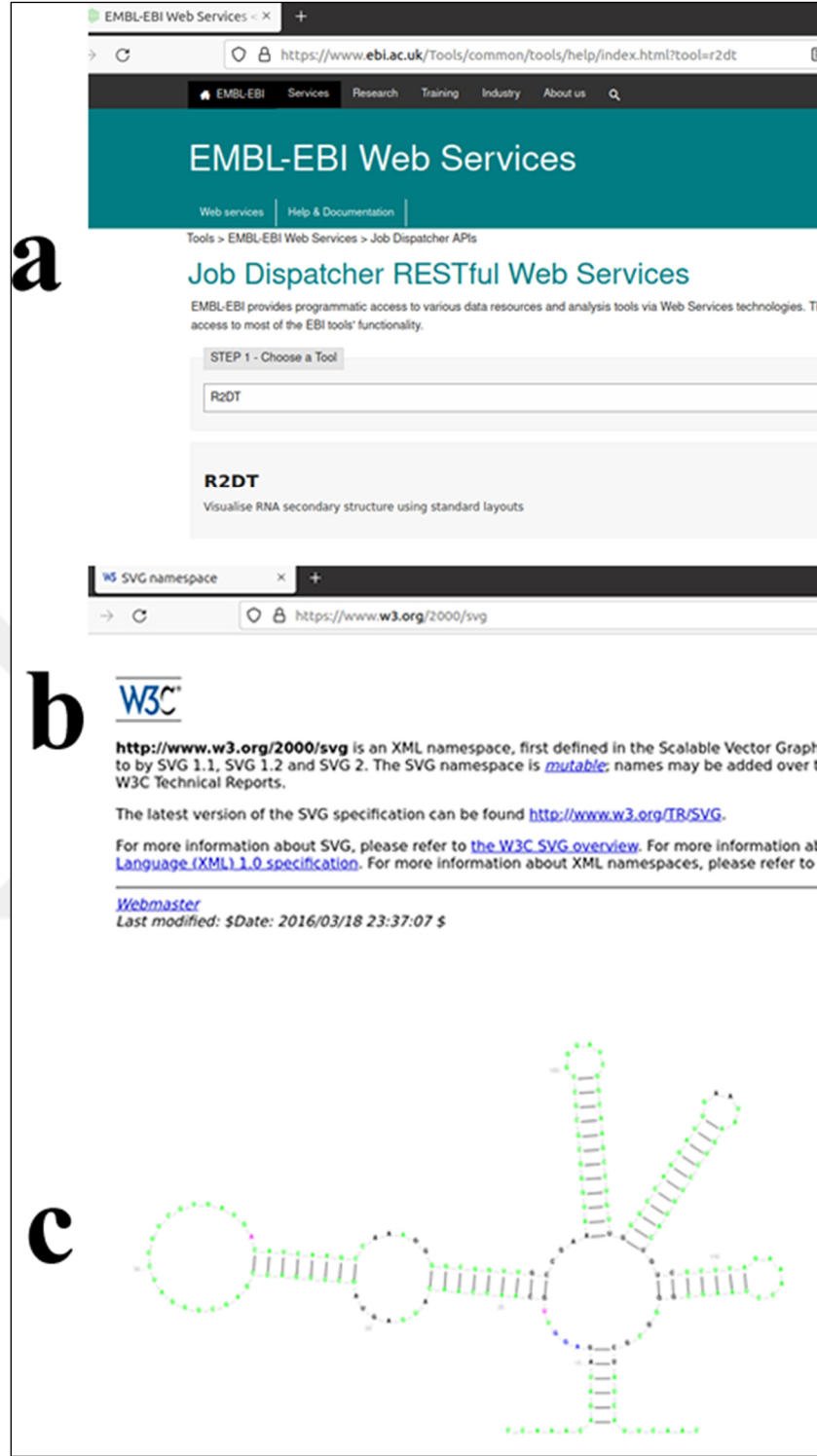
incelenebilir.

**b**

c

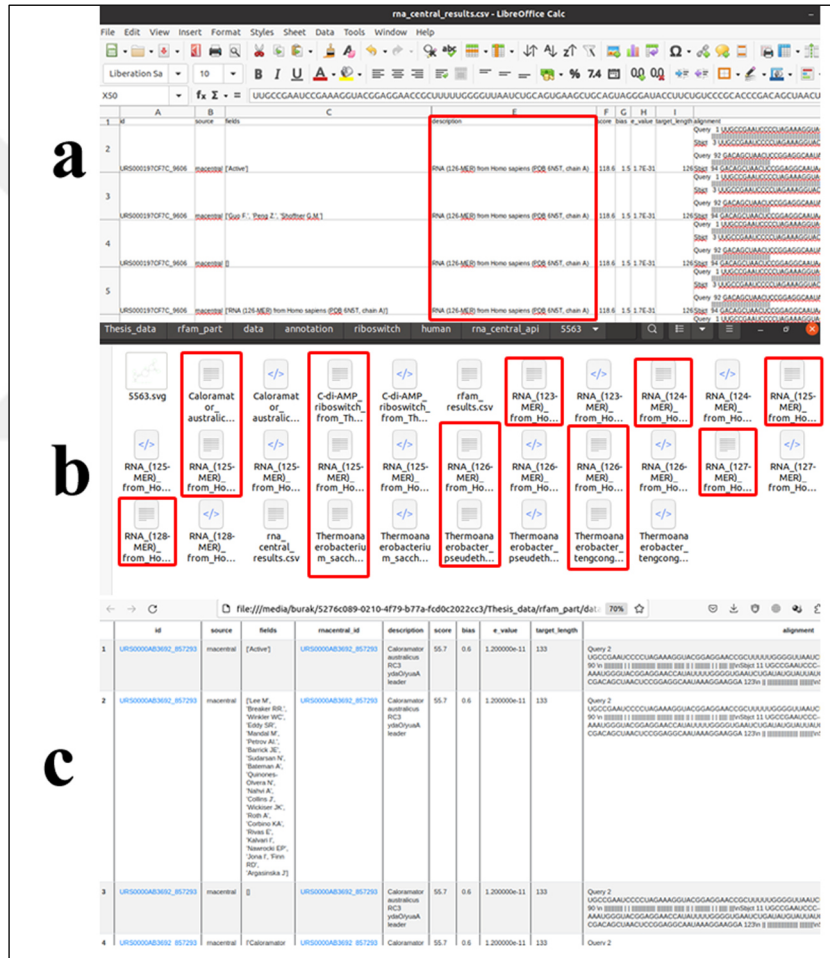
Üçüncü adım: Önceki adımda elde edilen “*rfam_results.csv*” sonuç dosyası içeriğinde “*pairwise alignment*” sonuçları yer almaktadır. Bu sonuçların biyoinformatik literatüründeki anlamlılığını ifade edebilmek için “*secondary structure prediction*” analizinin de yapılması bir gerekliliktir. Bu amaçla kod betiğine ait iş akışının bu adımında literatürde yer alan programlanabilir “*api*” kullanılmıştır (93). Böylelikle “*rfam_results.csv*” dosyası içinde tespit ettiğimiz Rfam biyolojik veri tabanı eşleşmelerine ait ikincil yapı tahmini ve “*svg*” formatındaki görsellerinin elde edilmesi işlemleri tamamlanmıştır. Bu adıma ait sürecin daha iyi anlaşılması için Resim 15 incelenebilir.





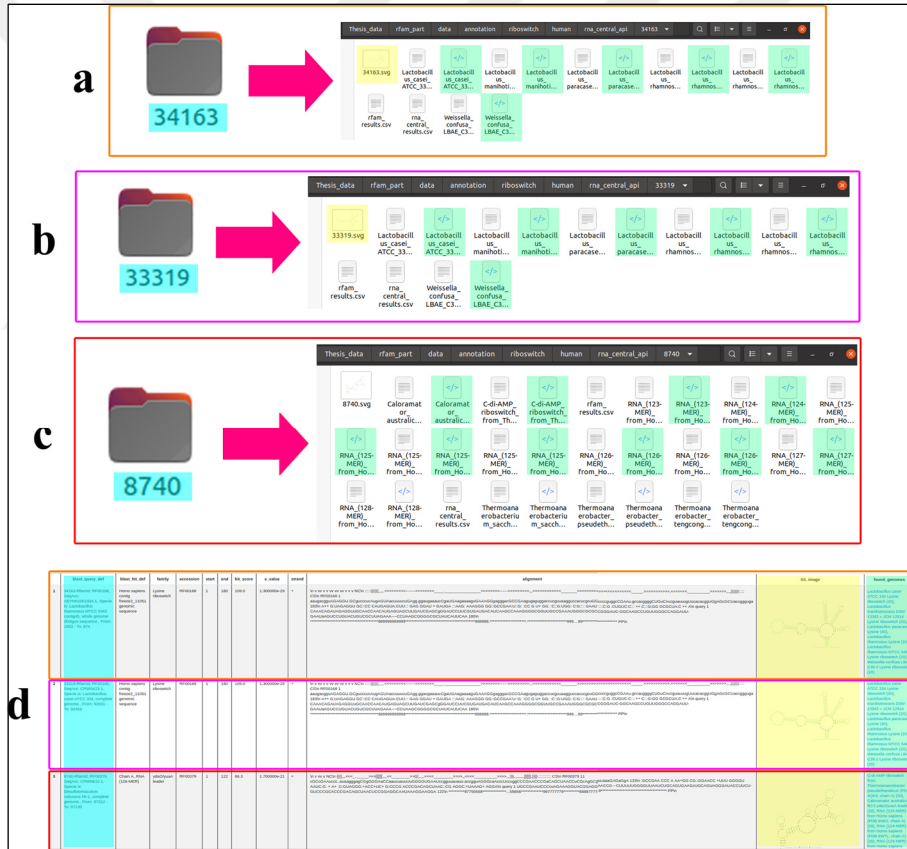
Resim 15. Riboşalter RNA ailesine ait üyeler için anotasyon analizi sürecinde kullanılan kod betiğine ait akış sürecinin üçüncü adımında yapılan işlemler. **(a)** İkincil yapı tahmini için kullanılan insan etkileşimine açık web aracının erişim sayfasından bir kesit. **(b)** Kullanılan programlanabilir “*api*” ile elde edilen çıktı formatını ait kaynak bilgi sayfasından bir kesit. **(c)** İkincil yapı analizi sonucunda elde edilen “*svg*” formatındaki çıktılarından örnek bir kesit.

Dördüncü adım: Bu kısımda diğer biyolojik veri tabanlarından elde edilen sonuçların filtrelenmesi ve genom adına göre ayrıştırılması işlemi yapılmıştır. Belirlenen eşik E-değerinin ($1e^{-10}$) altında kalan kayıtlar filtrelenmiş, diğer kayıtlar analiz dışında bırakılmıştır. Ardından anlamlı eşleşmelerin hangi genomlarda bulunduğu ve kaç adet olduğu tespit edilerek ana sonuç dosyası içerisinde her bir genome ait olanlar ayrı birer dosyaya önce “<genom_adi>.csv” formatında yazılmıştır. Son olarak da her bir genome ait “csv” dosyalarının “html” raporları oluşturulmuştur. Bu adıma ait sürecin daha iyi anlaşılması için Resim 16 incelenebilir.



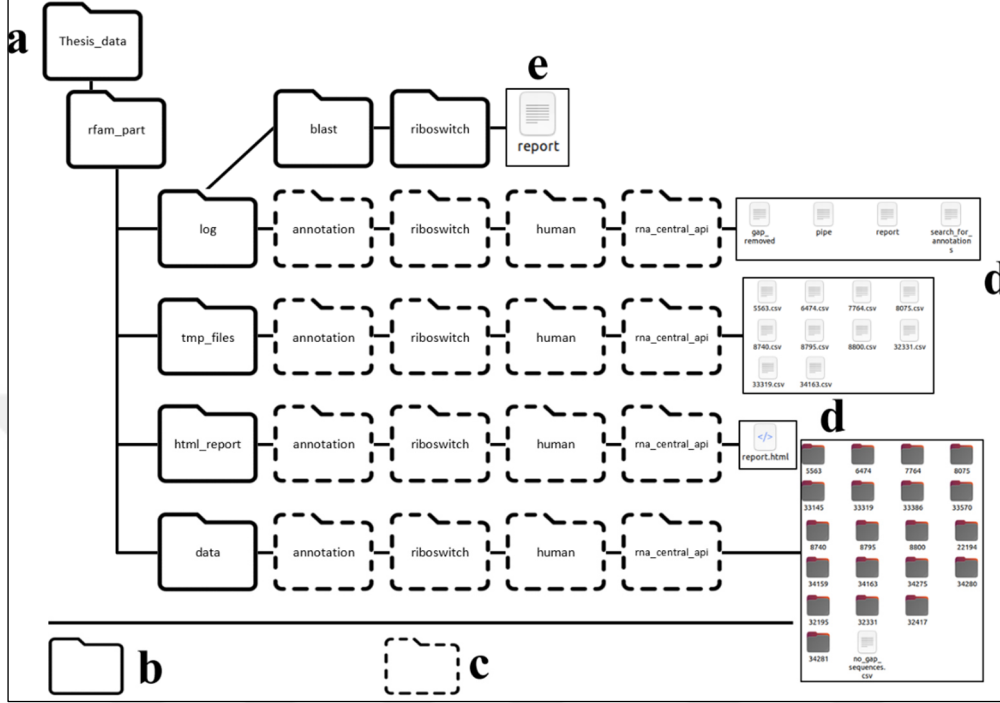
Resim 16. Riboşalter RNA ailesine ait üyeler için “anotasyon” analizi sürecinde kullanılan kod betiğine ait akış sürecinin dördüncü adımında yapılan işlemler. (a) Diğer biyolojik veri tabanlarında yapılan arama sonuçlarını içeren “rna_central_results.csv” dosyası içinde, genom adına göre ayrıştırma için kullanılan “description” sütununu gösteren bir kesit. (b) “description” sütunundaki anlamlı genom adlarına göre oluşturulan sonuç dosyalarından örnek bir kesit. (c) “description” sütunundaki anlamlı genom adlarına göre oluşturulan sonuç dosyalarının “html” formatındaki rapor dosyasına çevrilmiş bir örneğinden kesit.

Beşinci adım: Bu kısımda kod betiğine ait ana çıktı yani ana raporun oluşturulma işlemi gerçekleştirilmiştir. Rfam biyolojik veri tabanı içinde eşleşme bulunan ve her bir kayıt için kendi kayıt numarası ile oluşturulan klasör altında tutulan “*rfam_results.csv*” dosyasına ilk olarak ikincil yapı tahmin analizinden elde edilen “*svg*” görseli, yeni bir sütun olarak eklenmiştir. Ardından ilgili kaydın diğer veri tabanlarında bulunan eşleşmelerine ait veri sütuna da eklenmiştir. Bu veri sütunu içerisinde eşik E-değerinin altında kalan genom adı ve bu genomu ait kaç tane eşleşme bulunduğu bilgisi yer almaktadır. Ayrıca “*html*” formatında oluşturulan raporda bu veri sütunundan daha önce oluşturulan ilgili rapor dosyalarına da ulaşılabilir köprü bağlantısı da eklenmiştir. Bu adıma ait sürecin daha iyi anlaşılması için Resim 17 incelenebilir.



Resim 17. Riboşalter RNA ailesine ait üyeler için anotasyon analizi sürecinde kullanılan kod betiğine ait akış sürecinin beşinci adımında yapılan işlemler. (a) “*report.html*” adlı ana rapor dosyasının birinci satırında yer alan içeriğin dizindeki kaynağını gösteren bir kesit. (b) “*report.html*” adlı ana rapor dosyasının ikinci satırında yer alan içeriğin dizindeki kaynağını gösteren bir kesit (c) “*report.html*” adlı ana rapor dosyasının üçüncü satırında yer alan içeriğin dizindeki kaynağını gösteren bir kesit. (d) “*report.html*” adlı ana rapor dosyasının içeriğinden bir kesit.

Burada ifade edilen işlem adımlarının her biri için kullanılan dizin yapısını daha iyi anlamak için Şekil 8'i incelemeniz önerilir.



Şekil 8. Riboşalter RNA ailesine ait üyeler için “anotasyon” analizi süreci için tasarlanan dizin yapısı ve elde edilen dosyalar. (a) Kök dizini ifade eder. (b) Kod betiğinin çalışma sürecinde belirtilen kök dizin içerisinde daha önceki adımlarda oluşturulmuş klasörleri ifade eder. (c) Kod betiğinin çalışma sürecinde belirtilen kök dizin için ilk defa oluşturulan klasörleri ifade eder. (d) Kod betiği işlemini tamamladığında elde edilen çıktı, rapor ve log kayıt dosyalarını ifade eder. (e) Kod betiğinin çalışması için gerekli girdi dosyasını ifade eder.

3.3.7. İnsan (*H. sapiens*) Genomuna Ait GFF Dosyası Üzerinden Anotasyon Bölgelerinin Analizi

Çalışmanın bu kısmında NCBI biyolojik veri tabanında bulunan insan (*H. sapiens*) genomuna ait GFF dosyası *düzenli ifadeler* teknikleriyle analiz edilmiştir. İnsan genomunu oluşturan her bir kromozom bölgesinde anotasyon verisi bulunan kısımlar tespit edilmiş ve türleri belirlenmiştir. Ardından tespit edilen türlerin her bir kromozomda kaç adet yer aldığına dair veri tespit edilerek grafik çıktısı olarak verilmiştir. İleri bir analiz olarak “NCBI Genome Viewer Api” desteği kullanılarak insan genomunda henüz anotasyonu yapılmamış bölgeler tespit edilmiştir (94). Elde edilen tüm çıktılar “html” formatında sonuç rapor dosyası

olarak verilmiştir. Bu kısım için oluşturulan kod betiğinin gerçekleştirdiği işlem adımlarına ait akış süreci aşağıda ifade edilmiştir:

Birinci adım: Bu adımda NCBI biyolojik veri tabanında https://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/annotation/GRCh38_latest/refseq_identifiers/GRCh38_latest_genomic.gff.gz adresinde yer alan güncel GFF dosyası yerel diske indirilmiştir. Dosya “*gunzip*” formatında “*tar.gz*” uzantılı sıkıştırılmış bir dosyadır. Sıkıştırılmış sürümü yaklaşık 60 MB boyutunda, ayıklanmış sürümü yaklaşık 1.2 GB boyutundadır. İndirilen dosya, çalışma için yazarın kendi imkânlarıyla temin ettiği bilgisayarın hafızasına, tek seferde yüklenerek okunabilmesi için oldukça büyük boyutta olduğu için ayıklanmadan “*byte*” formatında parçalara ayrılarak okunmuştur. Bu adımda kod betiğinin ilgili kısmının yaptığı işlemleri daha iyi anlamak ve okunmuş GFF dosyasının içeriğine ait bir kesit görüntüyü incelemek için Resim 18’i incelemeniz önerilir.

a

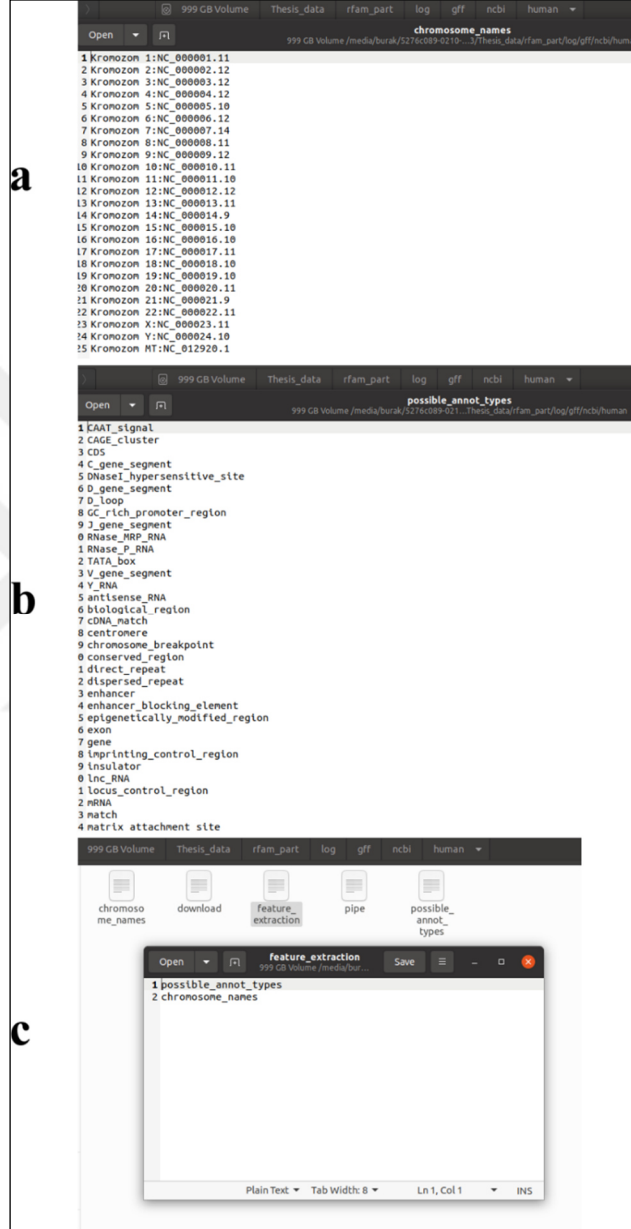
b

c

Resim 18. İnsan (*H. sapiens*) genomuna ait GFF dosyası üzerinden anotasyon bölgelerinin analizi sürecinde kullanılan kod betiğine ait akış sürecinin birinci adımında yapılan işlemler. **(a)** İnsan (*H. sapiens*) genomuna ait GFF dosyasının NCBI biyolojik veri tabanından indirilebilmesi için gerekli sayfa adresi ve içeriğinden bir kesit. **(b)** İnsan (*H. sapiens*) genomuna ait GFF dosyasının yerel diske indirilmiş konumundan bir kesit. **(c)** “Byte” formatında parçalarla ayrılarak okunan GFF dosya içeriğinden örnek bir kesit.

İkinci adım: Bu adımda parçalara ayrılarak okunan GFF dosyasının ilgili kısımlarından düzenli ifadeler (*regular expression*) komutları yardımıyla özellik çıkartma işlemleri yapılmıştır. Bu işlem için her bir GFF dosya parçası “Pandas” (95) iki boyutlu tablo formatına dönüştürülmüştür. Dönüştürülen tablodaki ilgili sütunlarda yer alan verilerden düzenli ifadeler (*regular expression*) yardımıyla sadece gerekli kısımlar çıkartılmıştır. Bu işlem her bir kromozoma ait dosya kısımları için tekrarlanarak, kromozom

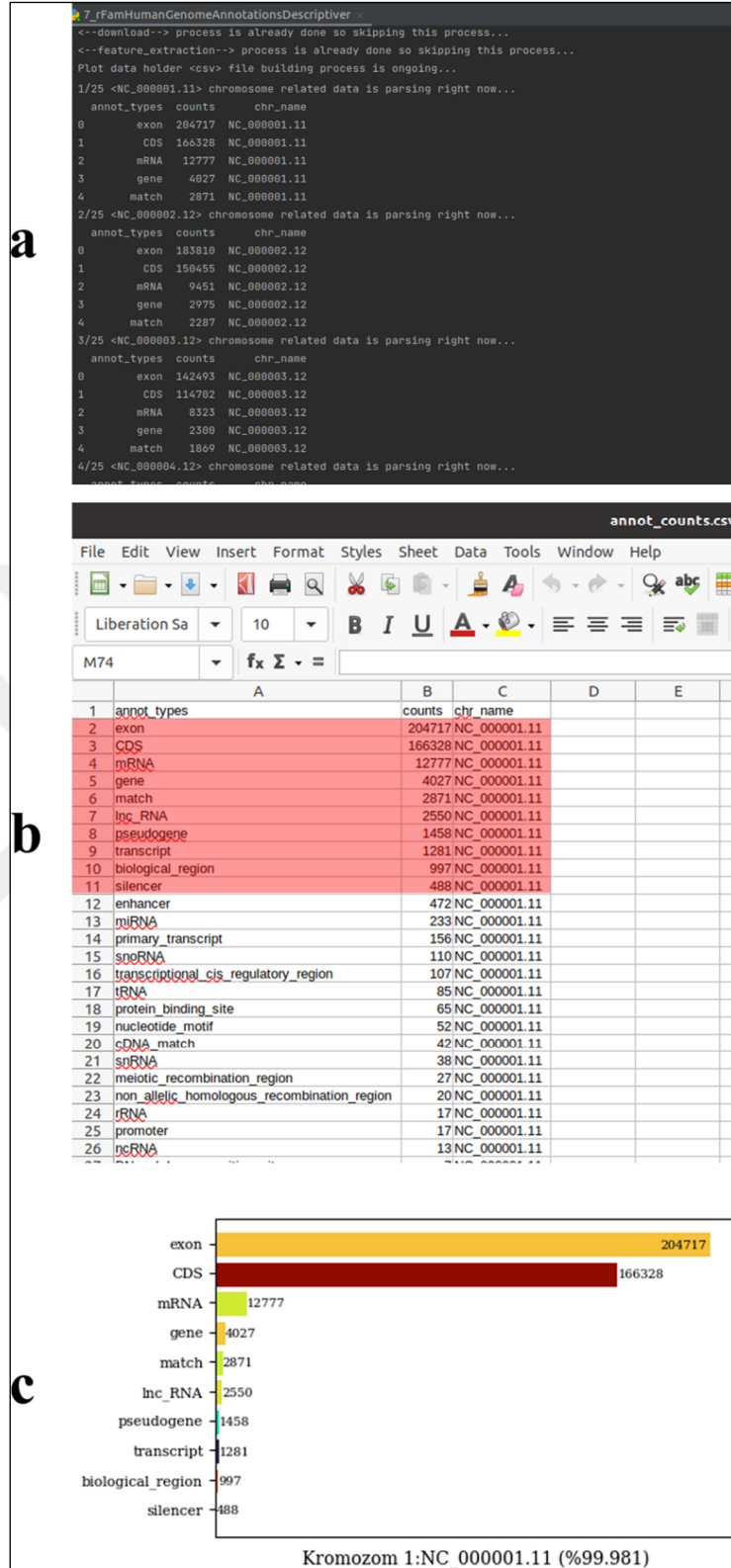
adları verisini içeren “*chromosome names*” ve anotasyon türleri verisini içeren “*possible_annot_types*” adlı “log” kayıt dosyalarına yazılmıştır. Bu adıma ait sürecin daha iyi anlaşılması için Resim 19 incelenebilir.



Resim 19. İnsan (*H. sapiens*) genomuna ait GFF dosyası üzerinden anotasyon bölgelerinin analizi sürecinde kullanılan kod betiğine ait akış sürecinin ikinci adımında yapılan işlemler. (a) İnsan (*H. sapiens*) genomuna ait GFF dosyasından kromozom isimleri verisinin çıkartılma işlemi sonucunda elde edilen “*chromosome names*” adlı dosya içeriğinden bir kesit. (b) İnsan (*H. sapiens*) genomuna ait GFF dosyasından anotasyon türleri verisinin çıkartılma işlemi sonucunda elde edilen “*possible_annot_types*” adlı dosya içeriğinden bir kesit. (c) Bu işlem adımı sonucunda elde edilen dosya adlarının saklandığı log kaydı dosya içeriğinden bir kesit.

Üçüncü adım: Bir önceki adımda çıkartılan özellik veri dosyaları kullanılarak her bir kromozom için tespit edilen anotasyon türlerinin frekans tablosu oluşturulmuştur. Bu dosya “*annot_counts.csv*” adıyla yerel diske kaydedilmiştir. Yerel diske kaydedilmiş olan bu dosya kullanılarak, her bir kromozom için elde edilmiş olan frekans tablolarını görselleştirmeye yarayan grafik çıktıları da elde edilmiştir. Elde edilen grafiklerde bütünlük olması için her bir kromozoma ait frekans tablolarındaki ilk on satır verisi kullanılmıştır. Bu adıma ait sürecin daha iyi anlaşılması için Resim 20 incelenebilir.





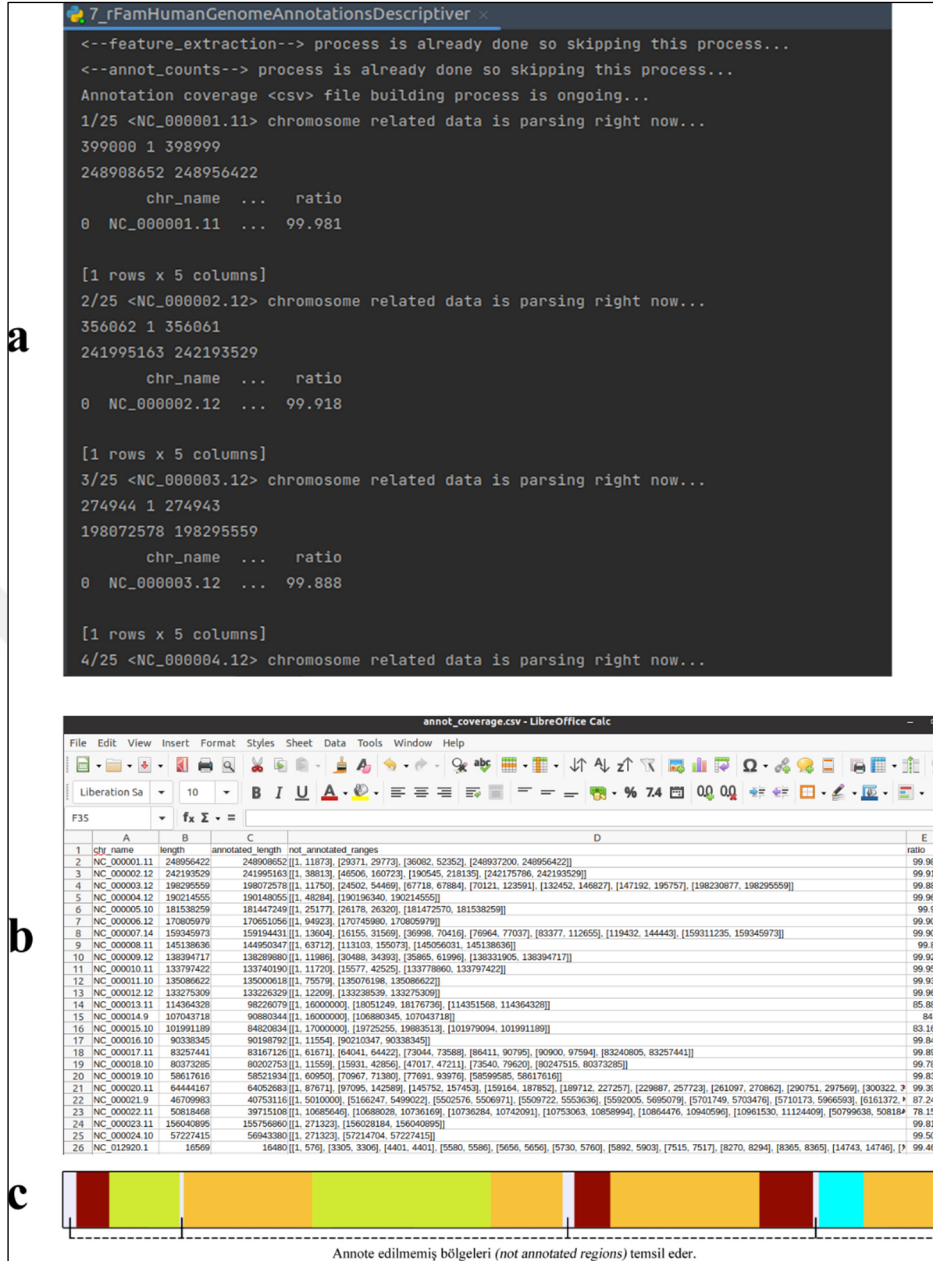
Resim 20. İnsan (*H. sapiens*) genomuna ait GFF dosyası üzerinden anotasyon bölgelerin analizi sürecinde kullanılan kod betiğine ait akış sürecinin üçüncü adımında yapılan işlemler. **(a)** Bu adım için kullanılan kod betiğinin çalışma sürecinden bir kesit. **(b)** “annot_counts.csv” dosya içeriğinden bir kesit. **(c)** Kromozom içerisinde yer alan anotasyon türlerinin dağılımı verisini tutan frekans tablolarının görselleştirilmesi sürecinde elde edilen örnek bir grafik çıktısı.

Dördüncü adım: GFF dosyası içeriğinde her bir kromozoma ait sekansın uzunluk verisi bulunmaktadır. Bahsi geçen dosya aynı zamanda, bu uzunluğun hangi kısımlarının anotasyon verisi içerdiğine ait veriyi de barındırmaktadır.

Bu adımda kod betiğinin ilgili kısmının yaptığı işlemin mantığı şöyle özetlenebilir. İlgili kromozomu bir sayı doğrusu gibi düşünelim. Bu sayı doğrusunun sınırı sıfırdan kromozomun uzunluğu kadar olsun. Her bir anotasyon bölgesi de bu sayı doğrusu içerisinde belirli bir alana karşılık gelecek şekilde temsil edilsin. Bazı bölgelerde çakışmaların olabileceği de kabul edilsin.

Yukarıdaki paragrafta verilen mantığa göre; burada yapılan işlem ise ilgili kromozoma ait alan içinde boş kalmış; yani henüz anotasyonu yapılmamış kısımların olup olmadığının tespit edilmesi ve eğer varsa bu kısımlara ait aralıkların “*annot_coverage.csv*” dosyası içine yazılmasıdır.

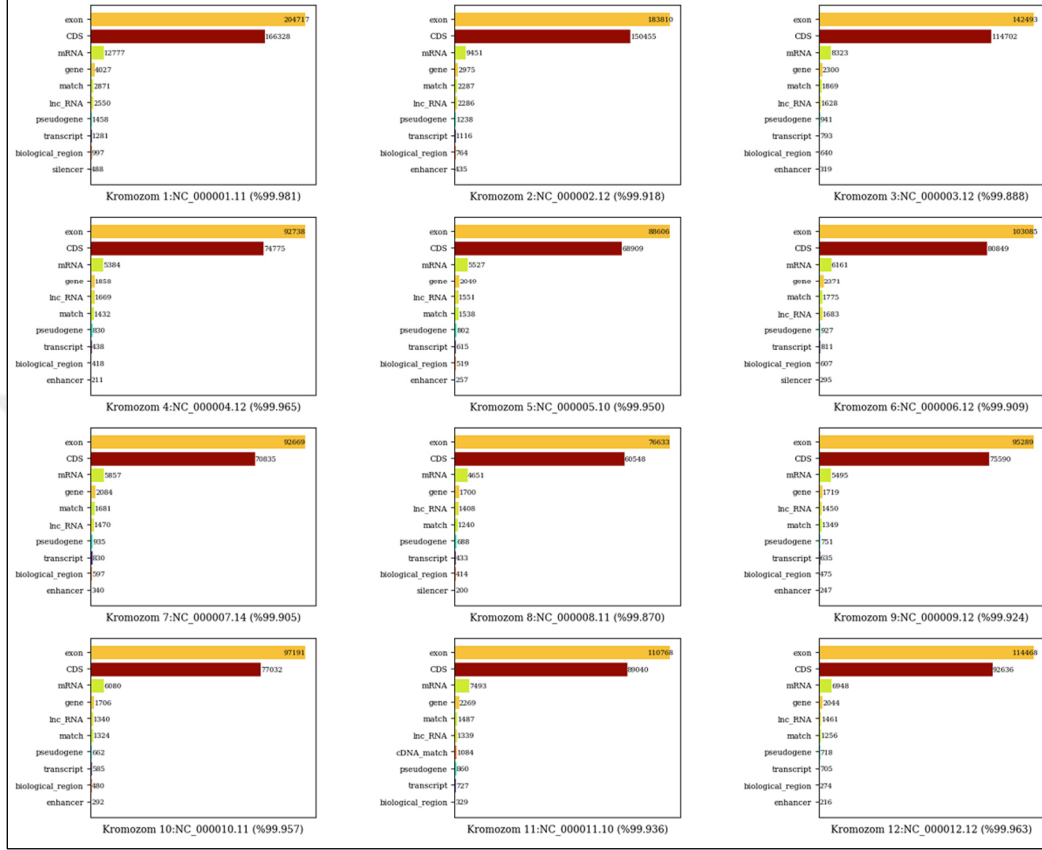
Özet olarak, bahsedilen bu mantıkla insan genomuna ait GFF dosyasındaki her bir kromozom için elde edilen veriler “*annot_coverage.csv*” dosyasına kaydedilmiştir. Bu veri kullanılarak her bir kromozomun yüzdelik oranda ne kadarlık kısmının anotasyonunun yapıldığına dair veri de aynı dosya içerisinde hesaplanmış ve kaydedilmiştir. Bu adıma ait sürecin daha iyi anlaşılması için Resim 21 incelenebilir.



Resim 21. İnsan (*H. sapiens*) genomuna ait GFF dosyası üzerinden anotasyon bölgelerinin analizi sürecinde kullanılan kod betiğine ait akış sürecinin dördüncü adımında yapılan işlemler. **(a)** Bu adım için kullanılan kod betiğinin çalışma sürecinden bir kesit. **(b)** “annot_coverage.csv” dosya içeriğinden bir kesit. **(c)** Anotasyonu yapılmamış bölgelerin tespit edilmesi sürecinde kullanılan mantığın betimlenmesinden bir kesit.

Beşinci adım: Bu adımda, girdi olarak bir önceki adımlarda oluşturulan “annot_counts.csv” ve “annot_coverage.csv” dosyaları kullanılmıştır. Bu dosyalar kullanılarak; insan genomunda yer alan her bir kromozoma ait anotasyon türü verilerinin frekans dağılım grafikleri yatay çubuk grafiği türünde oluşturulmuştur. Son olarak ise

oluşturulan tüm grafikler toplu halde tek bir grafik çıktısı olarak da elde edilmiştir. Bu adımda elde edilen çıktıları Resim 22 ve Resim 23 içerisinde yer verilmiştir.



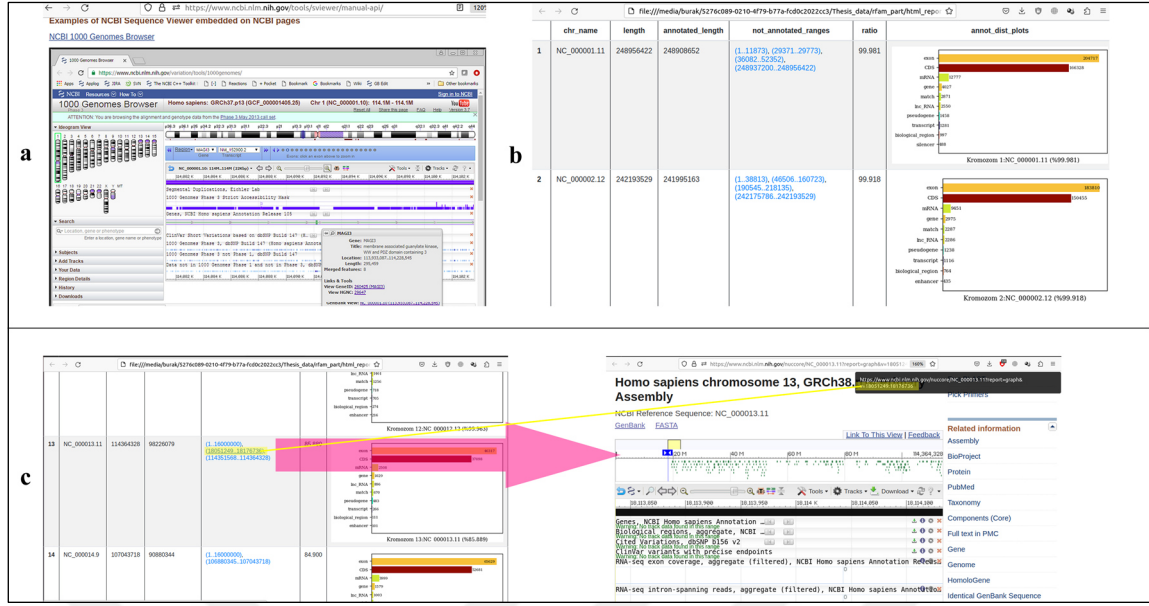
Resim 22. İnsan (*H. sapiens*) genomuna ait GFF dosyası üzerinden elde edilen anotasyon türleri frekans dağılım grafiklerin toplu olarak betimlenmesi birinci kısım.



Resim 23. İnsan (*H. sapiens*) genomuna ait *GFF* dosyası üzerinden elde edilen anotasyon türleri frekans dağılım grafiklerin toplu olarak betimlenmesi ikinci kısım.

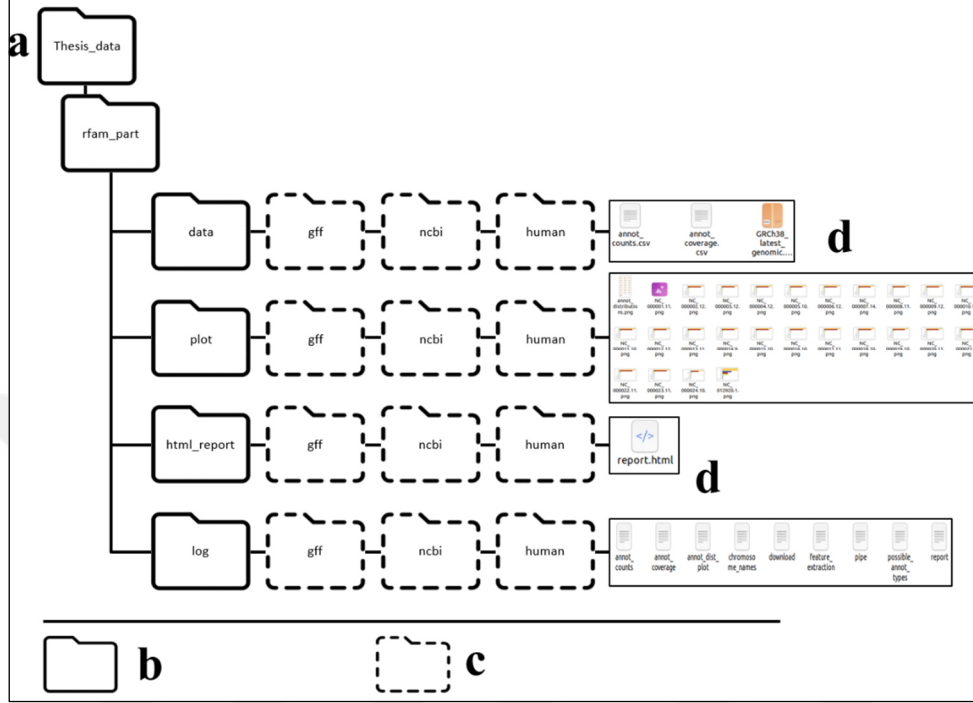
Altıncı adım: Kod betiğinin son bölümünü oluşturan bu kısımda her bir kromozoma ait tespit edilen anotasyonu yapılmamış kısımlar raporlanmış ve araştırmacıların bu kısımları daha iyi inceleyebilmesi için gerekli erişim bağlantısı verilmiştir (94). Bağlantı verilirken https://www.ncbi.nlm.nih.gov/tools/sviewer/url_access/ adresinde ifade edilen yönergeye uygun olarak parametreler seçilmiş ve ayarlanmıştır.

Her bir kromozoma ait yatay çubuk grafiği de rapor dosyası içerisinde konumlandırılmıştır. İçeriğinden bahsettiğimiz sonuç rapor dosyası “*html*” formatında dosya olarak diske yazılmıştır. Bu adıma ait sürecin daha iyi anlaşılması için Resim 24 incelenebilir.



Resim 24. İnsan (*H. sapiens*) genomuna ait GFF dosyası üzerinden anotasyon bölgelerinin analizi sürecinde kullanılan kod betiğine ait akış sürecinin altıncı adımında yapılan işlemler. **(a)** Araştırmacıların, *annotate* edilmemiş bölge olarak tespit ettiğimiz konumları görsel olarak inceleyebilmesi için kullanılan erişim bağlantısı içeriğinden bir kesit. **(b)** “*html*” formatındaki rapor dosyası içeriğinden bir kesit. **(c)** Annotate edilmemiş bölgelerin görsel olarak incelenebilmesi için kullanılan bağlantı parametreleri kullanılarak yapılan örnek bir erişim içeriğinden bir kesit

Burada ifade edilen işlem adımlarının her biri için kullanılan dizin yapısını daha iyi anlamak için Şekil 9’u incelemeniz önerilir.



Şekil 9. İnsan (*H. sapiens*) genomuna ait GFF dosyası üzerinden anotasyon bölgelerinin analizi süreci için tasarlanan dizin yapısı ve elde edilen dosyalar. **(a)** Kök dizini ifade eder. **(b)** Kod betiğinin çalışma sürecinde belirtilen kök dizin içerisinde daha önceki adımlarda oluşturulmuş klasörleri ifade eder. **(c)** Kod betiğinin çalışma sürecinde belirtilen kök dizin için ilk defa oluşturulan klasörleri ifade eder. **(d)** Kod betiği işlemini tamamladığında elde edilen çıktı, rapor, grafik ve log kayıtlarını ifade eder.

3.3.8. Rfam Biyolojik Veri Tabanında Bulunan Taksonomik Veri Setinin Derlenmesi ve “Sunburst” Grafik Çıktılarının Elde Edilmesi

Çalışmanın bu kısmında, ilgili kod betiği kullanılarak erişime açık SQL sunucusu üzerinden çekilen “sunburst.csv” dosyasında yer alan kayıtlar kullanılmıştır. Bahsi geçen bu kayıtlarda tutulan verilerden öncelikle araştırma konumuz olan Riboşalter RNA ailesine ait olan kayıtlar filtrelenmiştir. Filtrelenen kayıtlarda “json” formatındaki hiyerarşik yapıda düzenlenmiş veriler, çeşitli “python” paketlerinin kütüphanelerinde yer alan fonksiyonlar kullanılarak iki boyutlu veri tablosu formatına dönüştürülmüştür. Ardından dönüştürülen bu veriler “sunburst” grafiği oluşturmak için uygun girdi formatına çevrilmiş ve her bir Riboşalter aile üyesi “Rfam” elemanına ait grafik çıktıları düzenlenmiştir. Aynı veriler

kullanılarak her bir Riboşalter aile üyesi “*Rfam*” elemanının *annotate* edildiği kaynak genomların “*superkingdom*” taksonomi düzeyindeki frekans dağılım grafikleri de elde edilmiştir. Son olarak elde edilen tüm çıktılar araştırmacıların daha rahat inceleyebilmesi için “*html*” dosya formatında raporlanmıştır. Bu kısım için oluşturulan kod betiğinin gerçekleştirdiği işlem adımlarına ait akış süreci aşağıda ifade edilmiştir:

Birinci adım: Daha önceki adımlarda erişime açık “*Rfam*” SQL sunucusu kullanılarak “*Rfam*” biyolojik veri tabanının güncel sürümünün dosyaları yerel diske indirilmiştir. Bu kısımda indirilmiş olan dosyalardan “*sunburst.csv*” dosyası, bu başlık altında anlatılan işlemler için kullanılan kod betiğinin girdi dosyası olarak kullanılmıştır. Girdi dosyasındaki kayıtlar filtrelenerek sadece araştırma konumuz olan Riboşalter aile üyesi “*Rfam*” elemanlarına ait kayıtlar kullanılmıştır. Bu adıma ait sürecin daha iyi anlaşılması için Resim 25 incelenebilir.

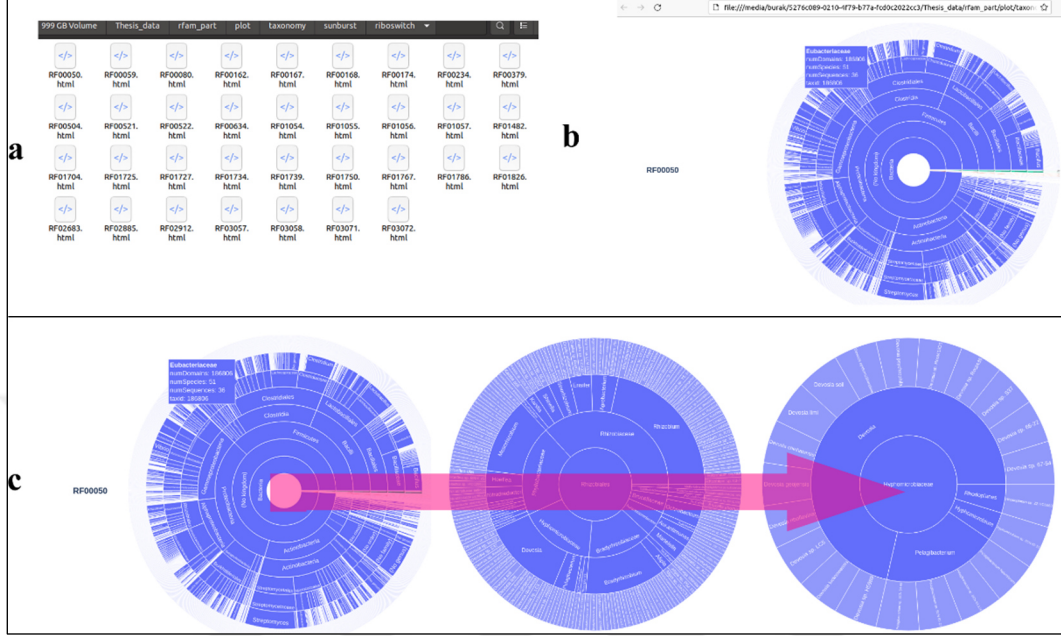


Resim 25. *Rfam* biyolojik veri tabanında bulunan taksonomik veri setinin derlenmesi ve “*sunburst*” grafik çıktıların elde edilmesi sürecinde kullanılan kod betiğine ait akış sürecinin birinci adımında yapılan işlemler. (a) Erişime açık *Rfam* SQL sunucusu kullanılarak *Rfam* biyolojik veri tabanının güncel sürümü üzerinden yerel diske indirilen “*sunburst.csv*” dosya konumunu gösteren bir kesit. (b) Filtreleme işlemi için kullanılan Riboşalter aile üyesi RNA elemanlarının, *Rfam* biyolojik veri tabanı özel erişim kimlik numaralarını gösteren dosya içeriğinden bir kesit. (c) Görsel çıktıların elde edilmesinde kullanılan “*sunburst.csv*” dosyasının içeriğinden bir kesit.

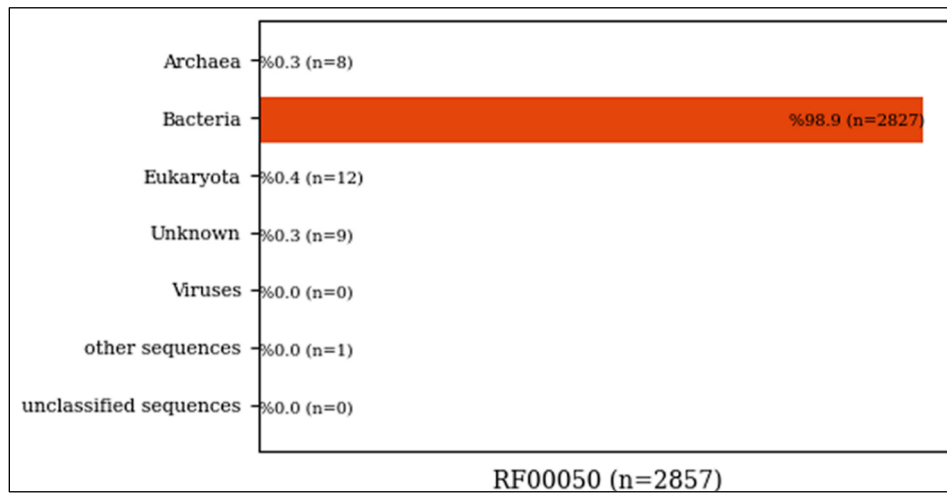
İkinci adım: Bu kısımda filtrelenmiş “sunburst.csv” dosyasında yer alan kayıtlar incelenmiştir. Kayıtların “byte” tipinde ve “json” veri formatı altında taksonomik hiyerarşiye göre düzenlenmiş olduğu görülmüştür. Öncelikle bu verinin daha okunabilir ve anlaşılabilir bir düzen olan satır ve sütunlar ile ifade edilen iki boyutlu veri tablosu düzenine çevrilmesinin bir gereklilik olduğu görülmüştür. Bu amaçla “Phyton” programlama dilinin “json” paketi altında yer alan aynı adlı kütüphanenin çeşitli fonksiyonları kullanılarak veri öncelikle düzleştirilmiştir. Düzleştirilerek hiyerarşik yapıdan kurtarılan veri içerisindeki taksonomik yapıya uygun örüntü keşfedilerek işlenmiş ve iki boyutlu tablo formatına başarıyla dönüştürülmüştür. Dönüştürülen dosyalar “<Rfam Özel Erişim Kimlik Numarası>_taxo.csv” (Örneğin, “RF00050_taxo.csv”) düzeninde adlandırılarak, çalışma kapsamına giren her bir RNA elamanı için oluşturulmuştur. Bu dosyalar kullanılarak çalışma kapsamına giren “Rfam” elemanlarının “annotate” edildiği kaynak genomların “superkingdom” taksonomi düzeyindeki frekans dağılım grafiklerinin oluşturulabilmesi için gerekli olan “all_members_faceted_bar.csv” dosyası da oluşturulmuştur.

Ardından “<Rfam Özel Erişim Kimlik Numarası>_taxo.csv” düzeninde kaydedilen dosyalar, “sunburst” grafiklerinin oluşturulabilmesi için gerekli girdi dosyası düzenine çevrilmiştir. Çevrimi yapılan dosyalar “<Rfam Özel Erişim Kimlik Numarası>_graph.csv” (Örneğin, “RF00050_graph.csv”) düzeninde adlandırılarak, çalışma kapsamına giren her bir RNA elamanı için oluşturulmuştur. Bu adıma ait sürecin daha iyi anlaşılması için Resim 26 incelenebilir.

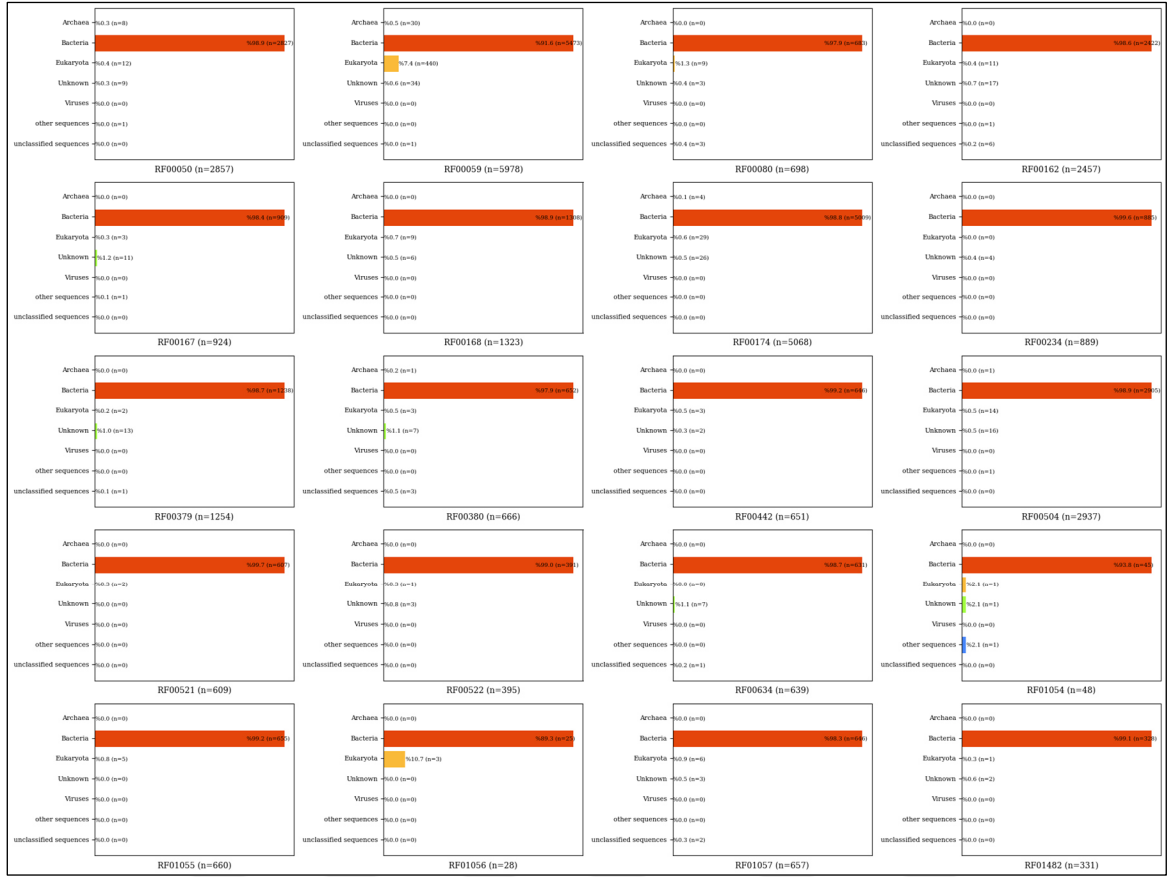
grafikleri ayrı ayrı ve toplu olarak elde edilmiştir. Bu adıma ait sürecin daha iyi anlaşılması için Resim 27-30 incelenebilir.



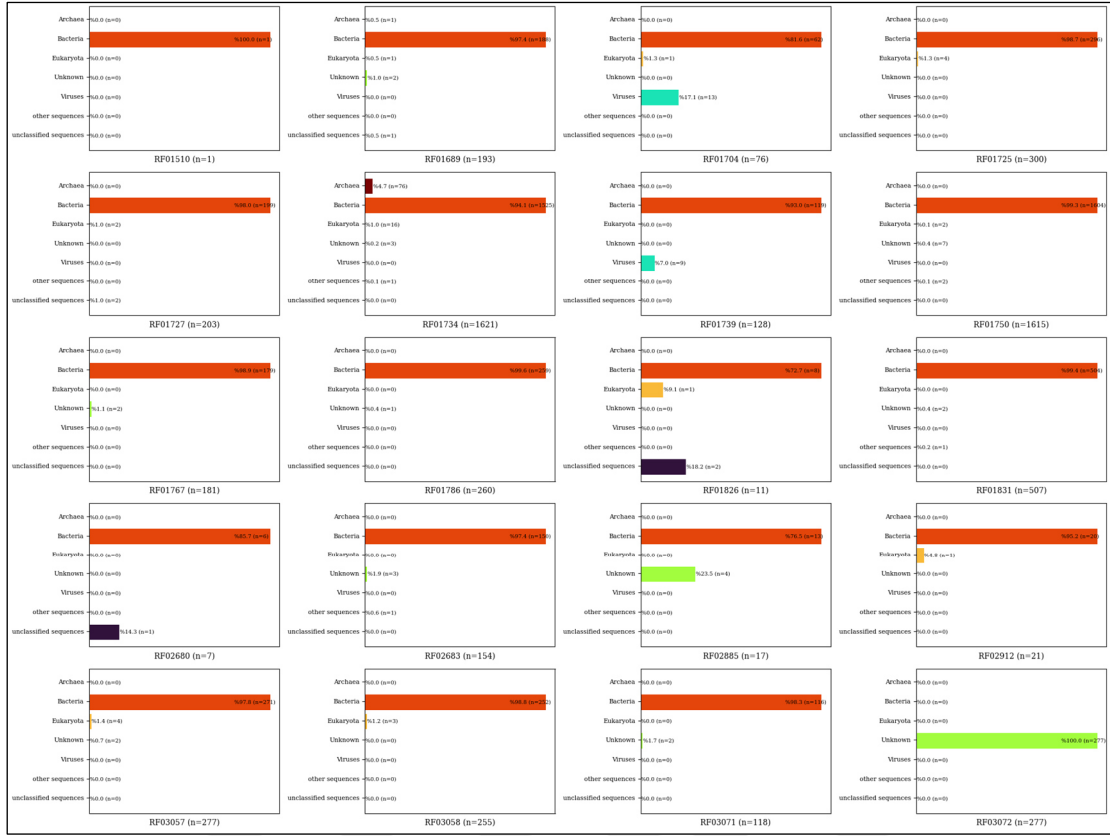
Resim 27. *Rfam* biyolojik veri tabanında bulunan taksonomik veri setinin derlenmesi ve “*sunburst*” grafik çıktılarının elde edilmesi sürecinde kullanılan kod betiğine ait akış sürecinin üçüncü adımında yapılan işlemler. (a) “*html*” formatında elde edilen “*sunburst*” grafik çıktılarının konumunu gösteren bir kesit. (b) Elde edilen “*sunburst*” grafik çıktı dosyası içeriğinden örnek bir kesit. (c) Kullanıcı etkileşimiyle “*sunburst*” grafik çıktısının alabileceği farklı görünümleri gösteren bir kesit.



Resim 28. “*all_members_faceted_bar.csv*” dosyası içerisinde yer alan her bir Riboşalter aile üyesi RNA elemanının, “*annotate*” edildiği kaynak genomların, “*superkingdom*” taksonomi düzeyindeki frekans dağılımının görselleştirilmesi sürecinde elde edilen örnek bir grafik çıktısı.

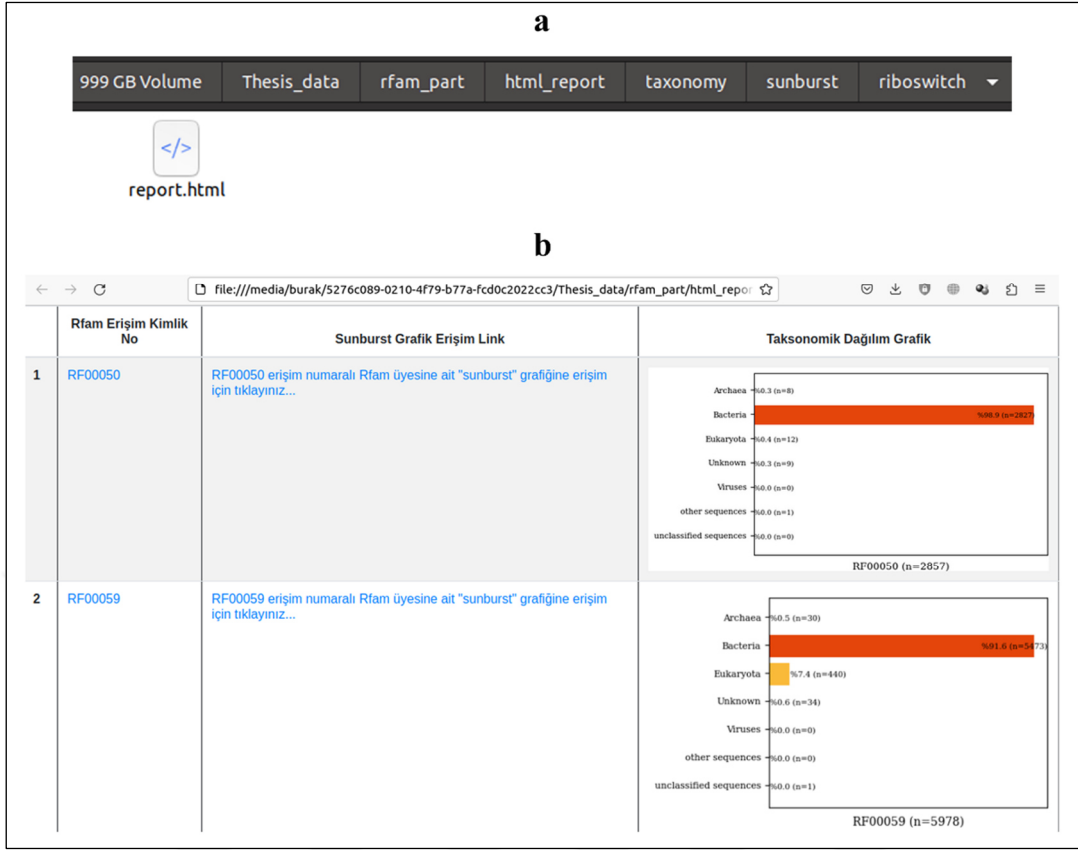


Resim 29. Riboşalter aile üyesi RNA elemanın, “*annotate*” edildiği kaynak genomların, “*superkingdom*” taksonomi düzeyindeki frekans dağılım grafiklerinin toplu olarak betimlenmesi birinci kısım.



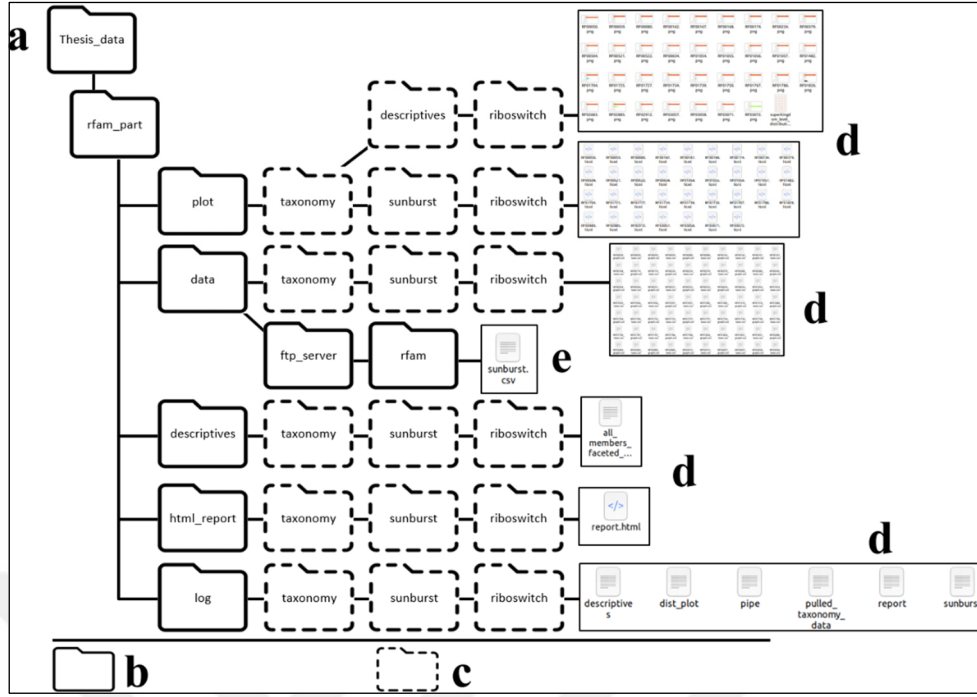
Resim 30. Riboşalter aile üyesi RNA elemanının, “*annotate*” edildiği kaynak genomların, “*superkingdom*” taksonomi düzeyindeki frekans dağılım grafiklerinin toplu olarak betimlenmesi ikinci kısım.

Dördüncü adım: Bu kısımda kod betiğine ait ana çıktı yani ana raporun oluşturulma işlemi gerçekleştirilmiştir. Rapor içeriğinde çalışma kapsamına giren her bir “*Rfam*” elemanı için oluşturulan “*sunburst*” grafiğine ait erişim bağlantısı, “*annotate*” edildiği kaynak genomların “*superkingdom*” taksonomi düzeyindeki frekans dağılım grafiği ve “*Rfam*” biyolojik veri tabanındaki içerik sayfasına erişim bağlantısı yer almaktadır. Rapor içeriğinin daha iyi anlaşılması için Resim 31 incelenebilir.



Resim 31. *Rfam* biyolojik veri tabanında bulunan taksonomik veri setinin derlenmesi ve “*sunburst*” grafik çıktılarının elde edilmesi sürecinde kullanılan kod betiğine ait akış sürecinin dördüncü adımında yapılan işlemler. **(a)** “*html*” formatındaki rapor dosyasının yerel diskteki konumundan bir kesit. **(b)** “*html*” formatındaki rapor dosyası içeriğinden bir kesit.

Burada ifade edilen işlem adımlarının her biri için kullanılan dizin yapısını daha iyi anlamak için Şekil 10’u incelemeniz önerilir.



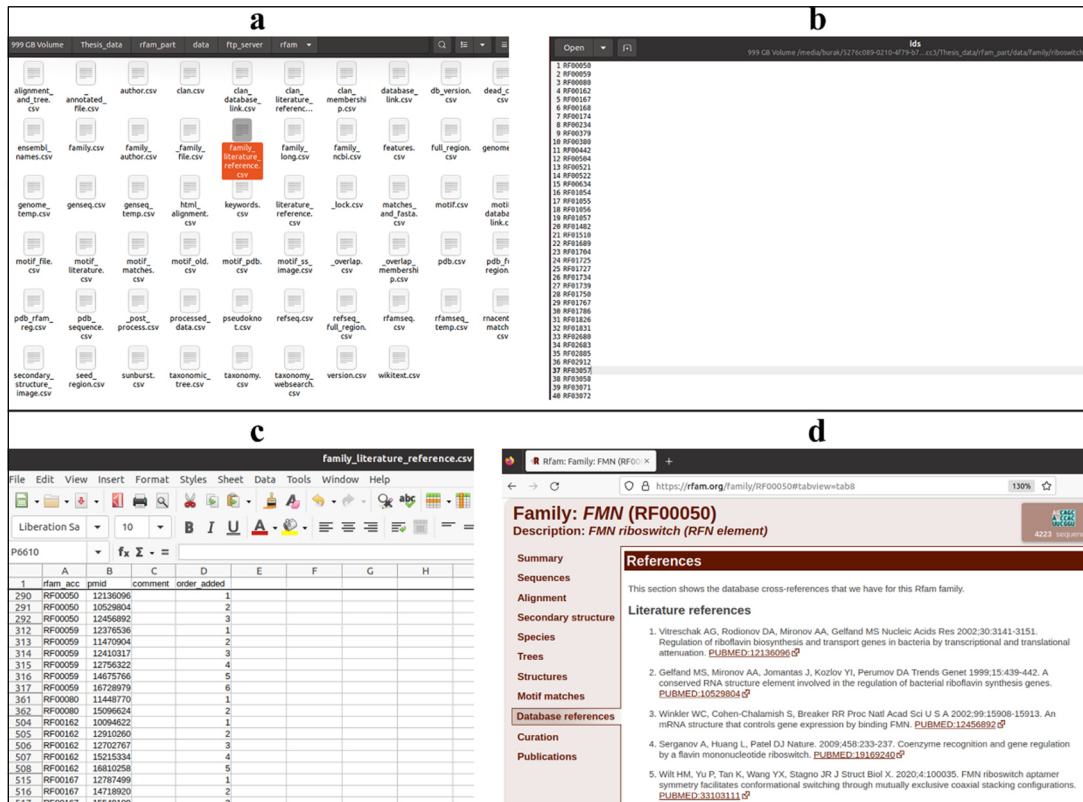
Şekil 10. Rfam biyolojik veri tabanında bulunan taksonomik veri setinin derlenmesi ve “sunburst” grafik çıktıların elde edilmesi süreci için tasarlanan dizin yapısı ve elde edilen dosyalar. (a) Kök dizini ifade eder. (b) Kod betiğinin çalışma sürecinde belirtilen kök dizin içerisinde daha önceki adımlarda oluşturulmuş klasörleri ifade eder. (c) Kod betiğinin çalışma sürecinde belirtilen kök dizin için ilk defa oluşturulan klasörleri ifade eder. (d) Kod betiği işlemini tamamladığında elde edilen çıktı, rapor, tanımlayıcı veri, grafik ve log kayıt dosyalarını ifade eder. (e) Kod betiğinin çalışması için gerekli girdi dosyasını ifade eder.

3.3.9. Rfam Biyolojik Veri Tabanında Bulunan Literatür Referansları Veri Seti Üzerinden Tanımlayıcı Grafik Çıktılarının Elde Edilmesi

Çalışmanın bu kısmında, ilgili kod betiği kullanılarak erişime açık SQL sunucusu üzerinden çekilen “family_literature_reference.csv” dosyasında yer alan kayıtlar kullanılmıştır. Bahsi geçen bu kayıtlarda tutulan verilerden, öncelikle araştırma konumuz olan Riboşalter RNA ailesine ait olan kayıtlar filtrelenmiştir. Ardından filtrelenen kayıtların içeriğindeki ilgili veri alanından, programlanabilir “api” desteği kullanılarak literatür referans veri seti oluşturulmuştur (96).

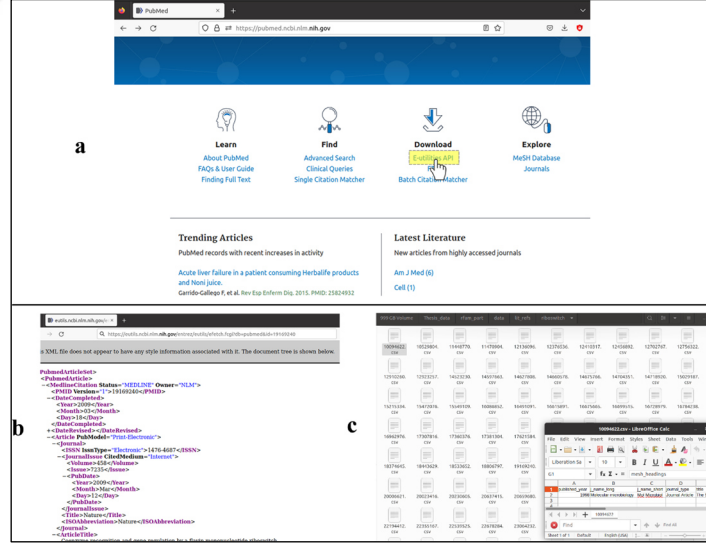
Elde edilmiş olan literatür referans veri seti içeriğinde çeşitli veri alanlarına ait veriler yer almaktadır. Bu veri alanları içerisinden araştırma kapsamında incelenmeye değer bulunanlara ait frekans dağılımı grafikleri de elde edilmiştir. Bu kısım için oluşturulan kod betiğinin gerçekleştirdiği işlem adımlarına ait akış süreci aşağıda ifade edilmiştir:

Birinci adım: Daha önceki adımlarda erişime açık “*Rfam*” SQL sunucusu kullanılarak “*Rfam*” biyolojik veri tabanının güncel sürümünün dosyaları yerel diske indirilmiştir. Bu kısımda indirilmiş olan dosyalardan “*family_literature_reference.csv*” dosyası, kod betiğinin girdi dosyası olarak kullanılmıştır. Girdi dosyasındaki kayıtlar filtrelenerek sadece araştırma konumuz olan Riboşalter aile üyesi “*Rfam*” elemanlarına ait kayıtlar kullanılmıştır. Bu adıma ait sürecin daha iyi anlaşılması için Resim 32 incelenebilir.



Resim 32. *Rfam* biyolojik veri tabanında bulunan literatür referansları veri seti üzerinden tanımlayıcı grafik çıktıların elde edilmesi sürecinde kullanılan kod betiğine ait akış sürecinin birinci adımında yapılan işlemler. (a) Kod betiğine girdi verisi sunan “*family_literature_reference.csv*” adlı dosyanın yerel diskteki konumundan bir kesit. (b) Girdi dosyasının filtrelenmesinde kullanılan dosya içeriğinden bir kesit. (c) Girdi dosyasının filtrelenmiş halinden bir kesit. (d) Girdi dosyasına ait içeriklerin, “*Rfam*” biyolojik veri tabanı sitesindeki karşılıklarının örnek görünümünden bir kesit.

İkinci adım: Filtrelenmiş “*family_literature_reference.csv*” dosya içeriği incelendiğinde, içerisindeki kayıtlarda her bir RNA elemanına ait “*pubmed*” referans numarası (PMID) verisi olduğu görülmektedir. PMID, NIH Ulusal Tıp Kütüphanesi tarafından *PubMed*’de indekslenen makalelere atanan bir numaradır (97). Açıklaması verilen bu referans numarası kullanılarak NCBI biyolojik tabanında yer alan programlanabilir “*api*” desteğinin kullandığımız kod betiğine bütünleştirmesi yapılarak girdi dosyasındaki her bir kayıt için gerekli veriler sunucuya istek olarak gönderilmiş ve sunucudan gelen yanıtlar beklenmiştir. Sunucudan gelen yanıt “*xml*” formatına göre düzenlenmiştir. Gelen yanıtların daha kolay ve anlaşılır şekilde derlenmesi için; “*xml*” türünden “*csv*” türündeki iki boyutlu tablo formatına dönüşüm işlemi için “*Phyton*” programlama dilinin kendi yerleşik kütüphanesinde bulunan “*ElementTree*” adlı kütüphanedeki ilgili fonksiyonlar kullanılmıştır. Bu süreç, her bir “*PMID*” kaydı için programlanabilir “*api*” yardımıyla oluşturan isteğe (*request*) karşı taraftaki sunucunun verdiği yanıt (*response*) “*<pubmed_indeksi>.csv*” formatında olacak şekilde yerel diske kaydedilmesiyle devam ettirilmiştir. Bu adıma ait sürecin daha iyi anlaşılması için Resim 33 incelenebilir.



Resim 33. *Rfam* biyolojik veri tabanında bulunan literatür referansları veri seti üzerinden tanımlayıcı grafik çıktıların elde edilmesi sürecinde kullanılan kod betiğine ait akış sürecinin ikinci adımında yapılan işlemler. (a) NCBI biyolojik veri tabanında yer alan ve “*Pubmed*” referans numaralı makaleler için kullanılan programlanabilir “*api*” desteği sunan internet sayfasından bir kesit. (b) Programlanabilir “*api*” desteği ile yapılan sorgulardan gelen “*xml*” formatındaki yanıtların içeriğinden örnek bir kesit. (c) Girdi dosyasının filtrelenmiş halinden bir kesit. (d) “*Api*” desteğiyle elde edilen ve “*<pubmed_indeksi>.csv*” formatında yerel diske kaydedilen dosyaların konumlarından ve içeriğinden bir kesit.

Üçüncü adım: Ardından elde edilen tüm “csv” dosyaları içerisinde ilgilendiğimiz veri alanlarına karşılık gelen “makalenin yayınlandığı derginin adı” ve “makalenin basım yılı” verileri “*published_year.csv*” ve “*j_name_short.csv*” adlı dosyalara toplu olarak kaydedilmiştir. Bu dosyalar içerisinde araştırma kapsamına giren Riboşalter RNA ailesine üye olan tüm RNA elemanlarına ait frekans dağılım verilerini içermektedir. Elde edilen dosyaların sonraki adımda oluşturulacak grafik çıktıları için düzenlenmesi gerekmektedir. Bu nedenle elde ettiğimiz dosyalarda bazı veri ön işleme adımları uygulanmıştır.

Şöyle ki; “*published_year.csv*” adlı dosya içeriğinde Riboşalter RNA ailesine üye olan tüm RNA elemanlarıyla ilgili yazılmış olan “*Pubmed*” referans numaralı makalelerin basıldığı yıllara göre frekans dağılımı verisi bulunmaktadır. Dosyadaki tüm veriler incelendiğinde bundan sonra kısaca “basım yılı” olarak ifade edeceğimiz veri alanının, 1998’den 2022’ye kadar yayılan veri aralığına sahip olduğu görülmüştür. Ancak her bir RNA elemanına ait kayıtlarda bu veri aralığının yalnızca küçük bir bölümüne karşılık gelecek veri dağılımının olduğu fark edilmiştir. Bu nedenle bahsi geçen veri alanından ortak bir grafik çıktı görünümü elde etmek için “basım yılı” verisi belirli yıl aralıklarını (örneğin 1998-2002, 2002-2007 vb.) kapsayacak şekilde gruplara ayrılarak kategorize edilmiştir. Bu şekilde veri ön işleme işlemine tabi tutulan dosya “*published_year_binned.csv*” adıyla yeni bir dosya olarak kaydedilmiştir. “Basım yılı” veri alanına karşılık gelen grafik çıktılarının elde edilmesinde veri ön işleme yapılmış olan “*published_year_binned.csv*” adlı dosya kullanılmıştır.

Diğer dosyamız ise “*j_name_short.csv*” adlı dosyadır. Bu dosya içeriğinde ise Riboşalter RNA ailesine üye olan tüm RNA elemanlarıyla ilgili yazılmış olan “*Pubmed*” referans numaralı makalelerin basıldığı dergilerin adlarına göre frekans dağılımı verisi bulunmaktadır. Bu dosya için ilgilendiğimiz veri alanı “makalenin yayınlandığı derginin adı” olarak isimlendirilmiştir ve bundan sonra kısaca “dergi adı” olarak ifade edilecektir. Dosyadaki tüm veriler incelendiğinde “dergi adı” olarak ifade ettiğimiz veri alanı içerisinde 34 farklı dergi adının yer aldığı görülmüştür. Ancak her bir RNA elemanına ait kayıtlara bakıldığında bu veri çeşitliliğinin yalnızca küçük bir miktarına karşılık gelecek veri dağılımının olduğu fark edilmiştir. Bu nedenle düzenli grafik çıktılarının elde edilmesi için; burada her bir kayıt önce içerdiği dergi sayısına göre büyükten küçüğe doğru sıralanmış ve ardından en yüksek frekansa sahip ilk on kayıt, analiz kapsamına alınmış; diğer veriler analiz

kapsamı dışında bırakılmıştır. Böylece her bir RNA elemanına ait grafik çıktısı elde edilirken standart düzen oluşturulmuştur.

Bu adıma ait sürecin daha iyi anlaşılması için Resim 34 incelenebilir.

a

published_year.csv - LibreOffice Calc

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	
A1	rfam_acc	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
2	RF00050	0	1	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
3	RF00059	0	0	1	2	2	0	0	4	2	1	0	1	0	0	1	1	0	0	0	0	0	0	0	0	
4	RF00080	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	
5	RF00162	1	0	0	0	2	1	0	1	0	0	0	3	0	2	1	0	0	1	0	0	0	0	1	0	
6	RF00187	0	0	0	1	2	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	
7	RF00188	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
8	RF00174	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
9	RF00234	0	0	0	0	1	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
10	RF00379	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
11	RF00380	0	0	0	0	0	1	0	0	1	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
12	RF00442	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	1	0	0	
13	RF00504	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	
14	RF00521	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
15	RF00522	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
16	RF00534	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
17	RF01054	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
18	RF01055	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	RF01056	0	0	0	0	0	0	2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	RF01057	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	RF01482	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0
22	RF01510	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	RF01689	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
24	RF01704	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
25	RF01725	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
26	RF01727	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

b

published_year_binned.csv

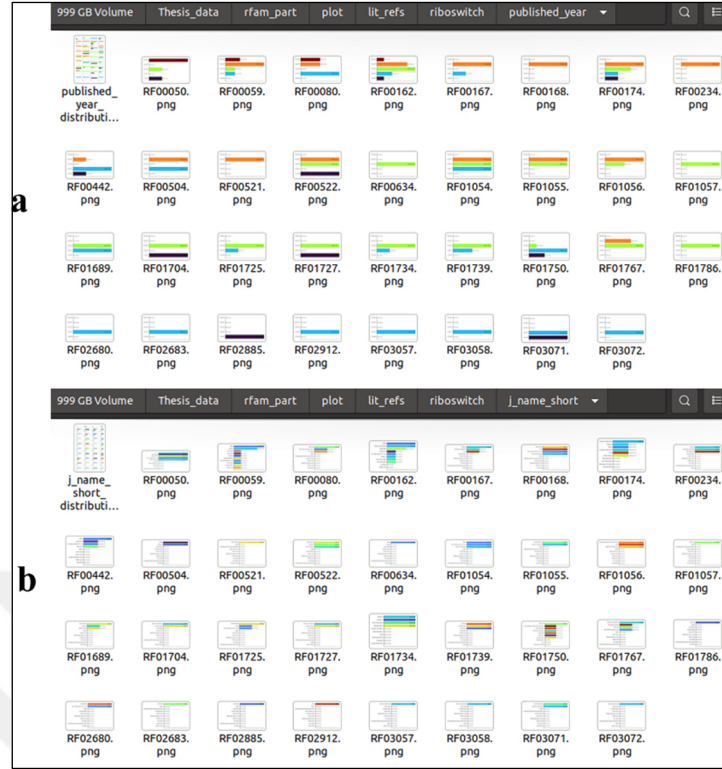
	A	B	C	D	E	F	G	H
A1	rfam_acc	1998-2002	2003-2007	2007-2012	2012-2017	2017-2022		
2	RF00050	3	0	1	0	1		
3	RF00059	3	8	2	2	0		
4	RF00080	1	1	0	2	0		
5	RF00162	1	4	5	2	1		
6	RF00187	0	3	0	1	0		
7	RF00188	0	4	0	0	0		
8	RF00174	0	3	2	1	1		
9	RF00234	0	3	0	0	0		
10	RF00379	0	1	0	0	0		
11	RF00380	0	2	3	0	0		
12	RF00442	0	1	0	3	1		
13	RF00504	0	1	0	1	0		
14	RF00521	0	1	0	0	0		
15	RF00522	0	1	1	0	1		
16	RF00534	0	0	1	0	0		
17	RF01054	0	1	1	0	0		
18	RF01055	0	1	1	0	0		
19	RF01056	0	2	1	0	0		
20	RF01057	0	0	1	0	0		
21	RF01482	0	0	2	3	0		
22	RF01510	0	0	1	0	0		
23	RF01689	0	0	2	2	0		
24	RF01704	0	0	1	0	1		
25	RF01725	0	0	3	1	0		
26	RF01727	0	0	1	0	1		

j_name_short.csv - LibreOffice Calc

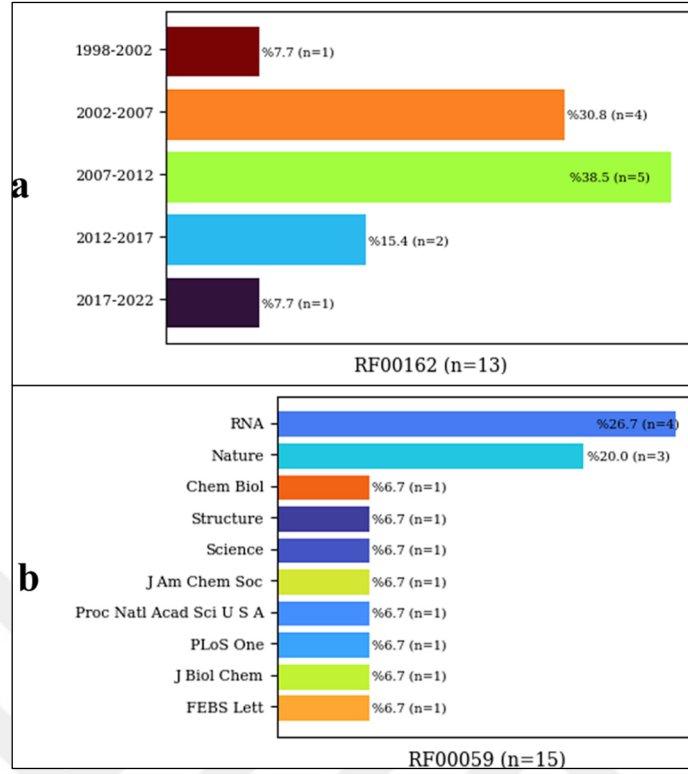
	A	B	C	D	E	F	G	H	I	J	K	L
A1	rfam_acc	Acta Crystallogr B	FEBS Lett	Genomics	Biochem Cell Biol	Biochemistry	Cell	Cell Chem Biol	Cell Rep	Chem Biol	Cold Spring Harb Symp Quant Biol	FEBS Lett
2	RF00050	0	0	0	0	0	0	0	0	0	0	0
3	RF00059	0	0	0	0	0	0	0	0	0	0	0
4	RF00080	0	0	0	0	0	0	0	0	0	0	0
5	RF00162	0	0	0	0	0	0	0	0	0	0	0
6	RF00187	0	0	0	0	0	1	0	0	0	0	0
7	RF00188	0	0	0	0	0	1	0	0	0	0	0
8	RF00174	0	0	0	0	0	1	0	0	0	0	0
9	RF00234	0	0	0	0	0	1	0	0	0	0	0
10	RF00379	0	0	0	0	0	0	0	0	0	0	0
11	RF00380	0	0	0	0	0	1	0	0	0	0	0
12	RF00442	0	0	0	0	0	0	0	0	0	0	0
13	RF00504	0	0	0	0	0	0	0	0	0	0	0
14	RF00521	0	0	0	0	0	0	0	0	0	0	0
15	RF00522	0	0	0	0	0	0	0	0	0	0	0
16	RF00534	0	0	0	0	0	0	0	0	0	0	0
17	RF01054	0	0	0	0	0	0	0	0	0	0	0
18	RF01055	0	0	0	0	0	0	0	0	0	0	0
19	RF01056	0	0	0	0	0	1	0	0	0	1	0
20	RF01057	0	0	0	0	0	0	0	0	0	0	0
21	RF01482	0	0	0	0	0	0	0	0	0	0	0
22	RF01510	0	0	0	0	0	0	0	0	0	0	0
23	RF01689	0	0	0	0	0	0	0	0	0	0	0
24	RF01704	0	0	0	0	0	0	0	0	0	0	0
25	RF01725	0	0	0	0	0	0	0	0	0	0	0
26	RF01727	0	0	0	0	0	0	0	0	0	0	0

Resim 34. Rfam biyolojik veri tabanında bulunan literatür referansları veri seti üzerinden tanımlayıcı grafik çıktılarının elde edilmesi sürecinde kullanılan kod betiğine ait akış sürecinin üçüncü adımında yapılan işlemler. (a) Makalelerin basım yılı verisini tutan “published_year.csv” adlı dosya içeriğinden bir kesit. (b) Makalelerin basım yılı verisinin veri ön işlemlerinden geçirilmiş halini içeren “published_year_binned.csv” adlı dosya içeriğinden bir kesit. (c) Veri ön işlemlerinden geçirilmiş, makalelerin yayınlandığı dergi isimleri verisini tutan, “j_name_short.csv” adlı dosya içeriğinden bir kesit.

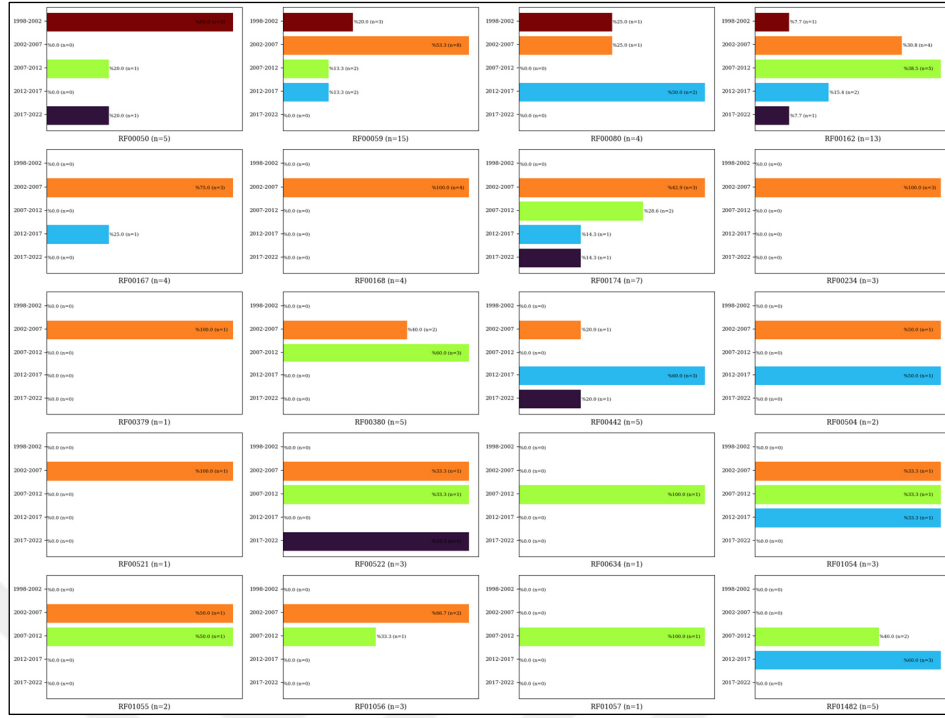
Dördüncü adım: Bu adımda kod betiğinin ilgili kısmı kullanılarak grafik çıktılarının elde edilme işlemleri gerçekleştirilmiştir. Kod betiği, üçüncü adımda açıklaması verilen veri ön işleme adımlarını başarıyla tamamlandıktan sonra ilgilendiğimiz iki veri türü için gerekli grafik çıktıları önce her bir RNA elemanı için ayrı ayrı olacak şekilde elde edilmiştir. Ardından ise tüm RNA elemanlarına ait grafik çıktılarının tek bir grafikte toplandığı grafik çıktıları da elde edilmiştir. Bu adıma ait sürecin daha iyi anlaşılması için Resim 35-40 incelenebilir.



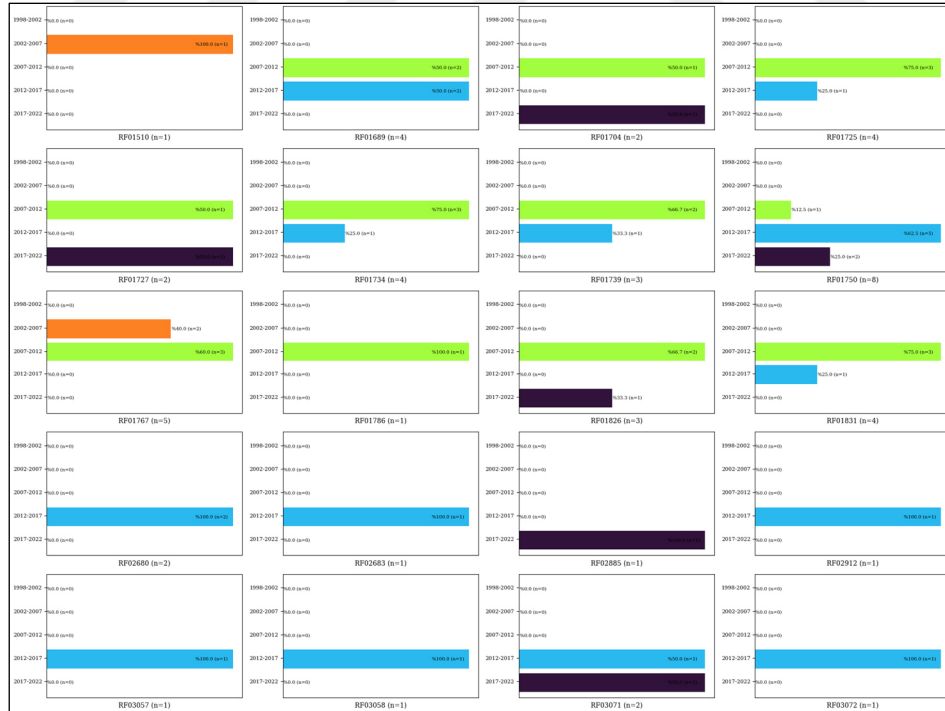
Resim 35. *Rfam* biyolojik veri tabanında bulunan literatür referansları veri seti üzerinden tanımlayıcı grafik çıktıların elde edilmesi sürecinde kullanılan kod betiğine ait akış sürecinin dördüncü adımında yapılan işlemler sonucunda elde edilen grafik çıktıları. **(a)** Makalelerin basım yılı verisinin RNA elemanlarına ait frekans dağılımlarını içeren grafik çıktıların yer aldığı yerel disk konumundan bir kesit. **(b)** Makalelerin yayınlandığı dergi isimleri verisinin RNA elemanlarına ait frekans dağılımlarını içeren grafik çıktıların yer aldığı yerel disk konumundan bir kesit.



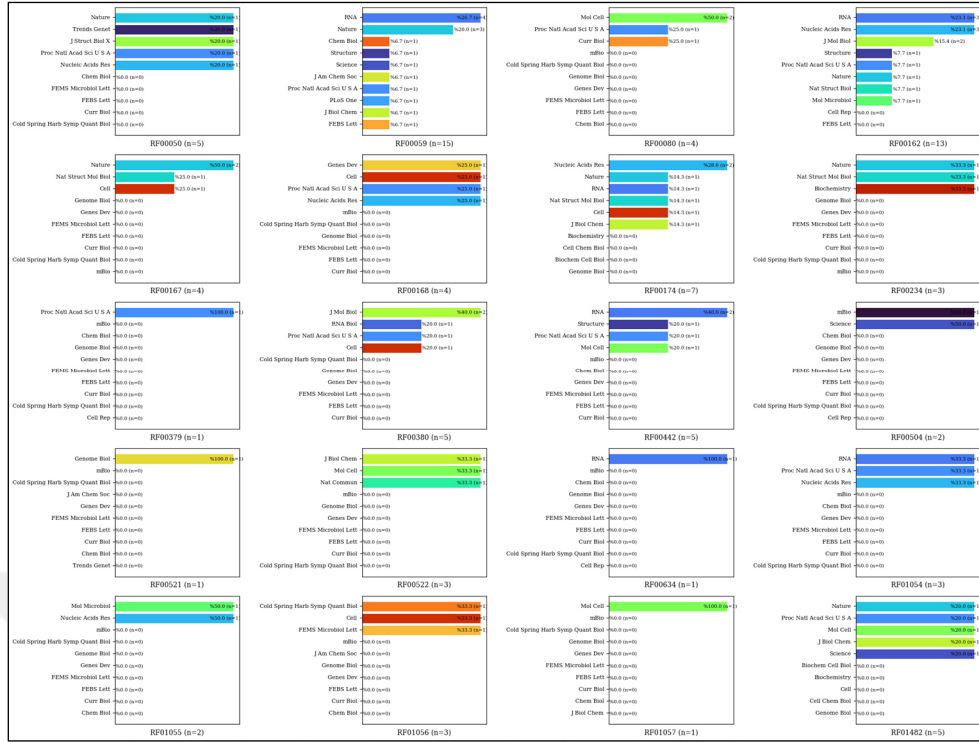
Resim 36. *Rfam* biyolojik veri tabanında bulunan literatür referansları veri seti üzerinden tanımlayıcı grafik çıktıların elde edilmesi sürecinde kullanılan kod betiğine ait akış sürecinin dördüncü adımında yapılan işlemler sonucunda elde edilen grafik çıktıları. **(a)** Makalelerin basım yılı verisinin RNA elemanlarına ait frekans dağılımlarını içeren grafik çıktıların tekil gösteriminden örnek bir kesit. **(b)** Makalelerin yayınlandığı dergi isimleri verisinin RNA elemanlarına ait frekans dağılımlarını içeren grafik çıktıların tekil gösteriminden örnek bir kesit.



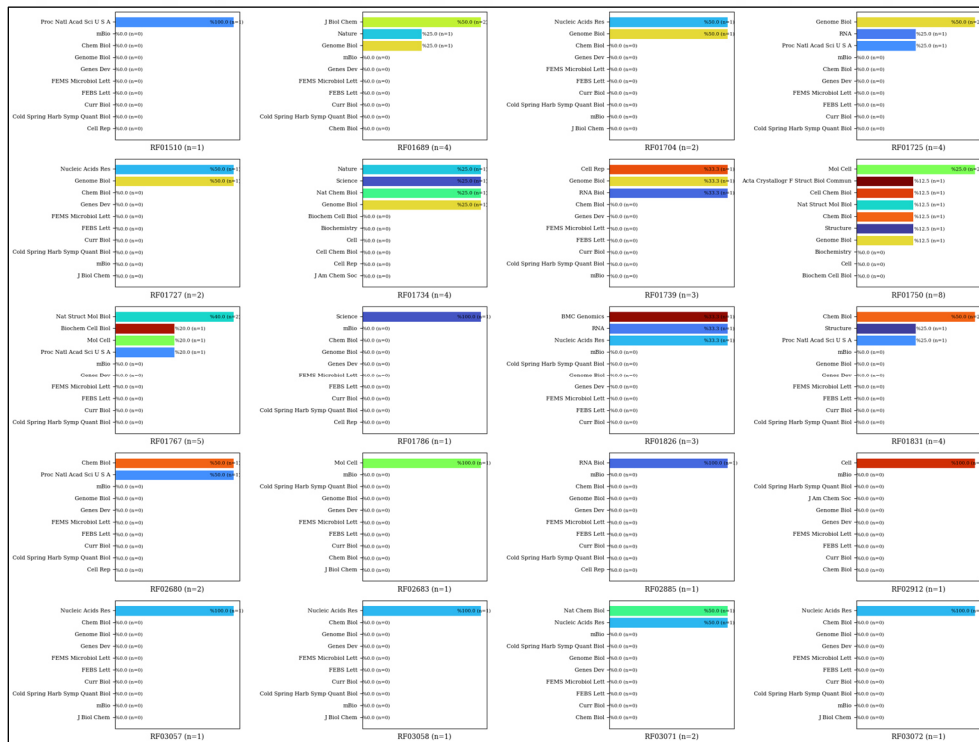
Resim 37. Makalelerin basım yılı verisinin RNA elemanlarına ait frekans dağılımlarını içeren grafik çıktılarından toplu gösteriminin birinci kısmından alınan kesit.



Resim 38. Makalelerin basım yılı verisinin RNA elemanlarına ait frekans dağılımlarını içeren grafik çıktılarından toplu gösteriminin ikinci kısmından alınan kesit.

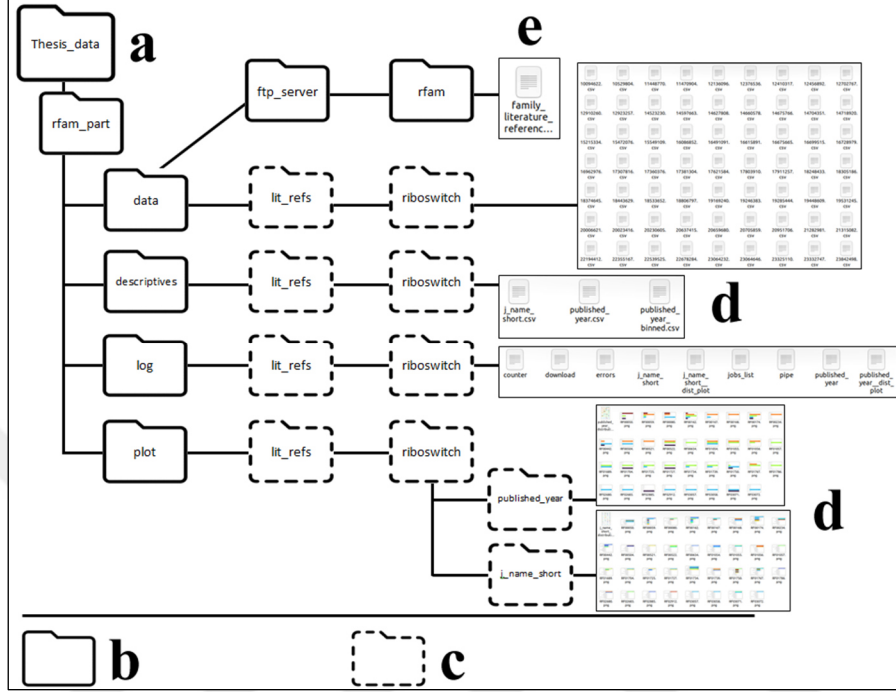


Resim 39. Makalelerin yayınlandığı dergi isimleri verisinin RNA elemanlarına ait frekans dağılımlarını içeren grafik çıktılarının toplu gösteriminin birinci kısmından alınan kesit.



Resim 40. Makalelerin yayınlandığı dergi isimleri verisinin RNA elemanlarına ait frekans dağılımlarını içeren grafik çıktılarının toplu gösteriminin ikinci kısmından alınan kesit.

Burada ifade edilen işlem adımlarının her biri için kullanılan dizin yapısını daha iyi anlamak için Şekil 11'i incelemeniz önerilir.



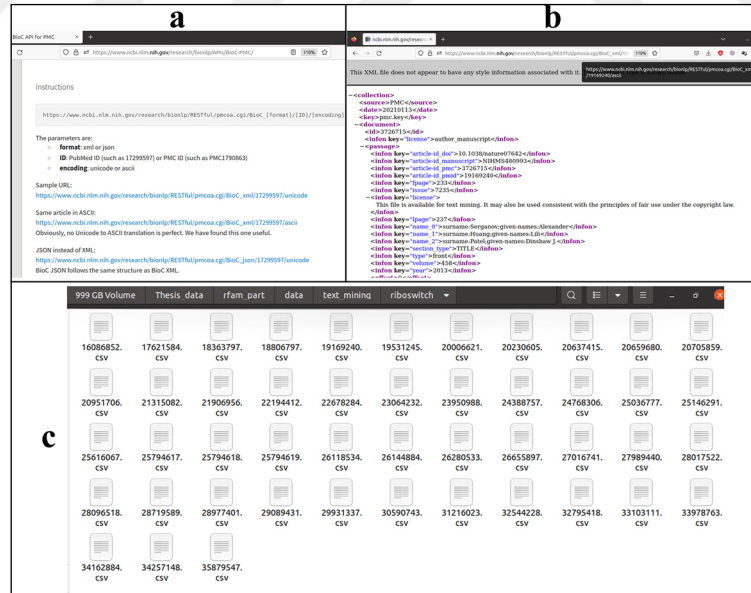
Şekil 11. Rfam biyolojik veri tabanında bulunan literatür referansları veri seti üzerinden tanımlayıcı grafik çıktıların elde edilmesi süreci için tasarlanan dizin yapısı ve elde edilen dosyalar. (a) Kök dizini ifade eder. (b) Kod betiğinin çalışma sürecinde belirtilen kök dizin içerisinde daha önceki adımlarda oluşturulmuş klasörleri ifade eder. (c) Kod betiğinin çalışma sürecinde belirtilen kök dizin için ilk defa oluşturulan klasörleri ifade eder. (d) Kod betiği işlemini tamamladığında elde edilen çıktı, tanımlayıcı veri, grafik ve log kayıt dosyalarını ifade eder. (e) Kod betiğinin çalışması için gerekli girdi dosyasını ifade eder.

3.3.10. Rfam Biyolojik Veri Tabanında Bulunan Literatür Referansları Veri Seti Üzerine Metin Madenciliği Uygulanarak Grafik Çıktılarının Elde Edilmesi

Bir önceki konu başlığında yapılan işlemlere ilave olarak bu kısımda, NCBI biyolojik tabanında yer alan ve “Pubmed” referans numaralı makaleler üzerinde metin madenciliği işlemleri yapabilmek için düzenlenmiş verilere erişimi sağlayan, programlanabilir “api” desteği kullanılmıştır. NCBI biyolojik tabanında yer alan ve metin madenciliği analizlerinde kullanılması için özel olarak düzenlenen “Pubmed” referans numaralı makale havuzu gelişim aşamasındadır ve bazı eksiklikleri bulunmaktadır. Özellikle dijital dönüşümü henüz yapılamamış olan eski tarihli makaleler için gerekli veriler eksiktir. Bu nedenle bizim girdi dosyamızda 40 farklı RNA elemanına ait “pubmed” referans numaralı kayıtlardan 30

tanecini kapsayan veriye ulaşılmıştır. Bundan dolayı analizler 30 RNA elemanına ait veriler üzerinden yapılmış ve elde edilen tüm çıktılar “*html*” formatında raporlanmıştır. Bu kısım için oluşturulan kod betiğinin gerçekleştirdiği işlem adımlarına ait akış süreci aşağıda ifade edilmiştir:

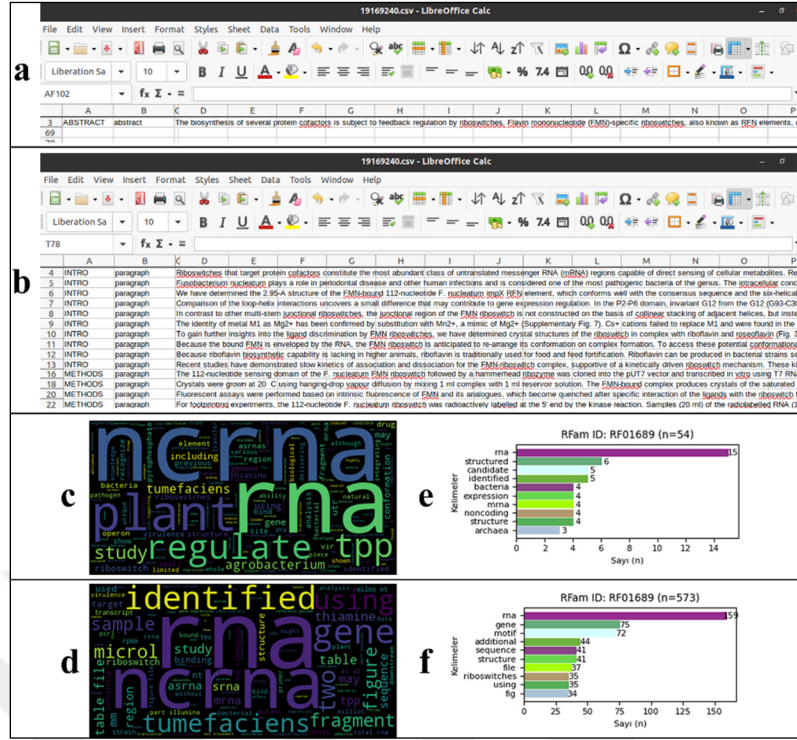
Birinci adım: Bir önceki konu başlığına benzer şekilde bu kısımda da “*family_literature_reference.csv*” dosyasının filtrelenmiş hali girdi olarak kullanılmıştır. Dosya kayıtları içerisinde bulunan “*Pubmed*” referans numarası <https://www.ncbi.nlm.nih.gov/research/bionlp/APIs/BioC-PMC/> adresinde yer alan metin madenciliği web servisine sorgu olarak gönderilmiş ve yanıt alınan *pubmed* referans numarası içerikleri <*pubmed_referans_no*>.csv düzeninde yerel diske kaydedilmiştir (98). Gönderilen sorgulara sunucudan gelen yanıtlar incelenmiştir. Gelen yanıtların, sorgu olarak gönderilen “*Pubmed*” referans numaralı makalenin içerik türüne göre (özet, makale metni, referans vb.) gruplanarak verildiği görülmüştür. Ayrıca her biri gruba uyan metin içeriğinin iki boyutlu tablo düzeninde derlenerek yerel diske yazıldığı anlaşılmıştır. Bu adıma ait sürecin daha iyi anlaşılması için Resim 41 incelenebilir.



Resim 41. Rfam biyolojik veri tabanında bulunan literatür referansları veri seti üzerine metin madenciliği uygulanarak grafik çıktıların elde edilmesi sürecinde kullanılan kod betiğine ait akış sürecinin ikinci adımında yapılan işlemler. (a) NCBI biyolojik veri tabanında yer alan ve “*Pubmed*” referans numaralı makalelerde metin madenciliği işlemleri için oluşturulmuş, programlanabilir “*api*” desteği sunan internet sayfasından bir kesit (b) Programlanabilir “*api*” desteği ile yapılan sorgulardan gelen “*xml*” formatındaki yanıtların içeriğinden örnek bir kesit. (c) “*Api*” desteğiyle elde edilen ve “<*pubmed_referans_no*>.csv” formatında yerel diske kaydedilen dosyaların konumundan bir kesit.

İkinci adım: İki boyutlu tablo düzenine göre elde edilen içerik, metin madenciliği işlemleri için kullanılan çeşitli fonksiyonların girdi formatına dönüştürülebilmek için oldukça kullanışlı bir tasarıma sahiptir. Burada yapılan işlemlerdeki mantığı şöyle ifade edebiliriz. Elde edilen makale içerikleri iki şekilde analiz edilmiştir. Birinci analizde sadece makale özeti için yazılmış metin içeriklerini analiz ederken; ikinci analiz özet haricindeki tüm makale metnini (giriş, materyal ve metot, bulgular, tartışma ve sonuç) analiz etmektedir. Başlıklar, referans ve tablo ile grafik açıklama metinleri ise analiz kapsamı dışında tutulmuştur.

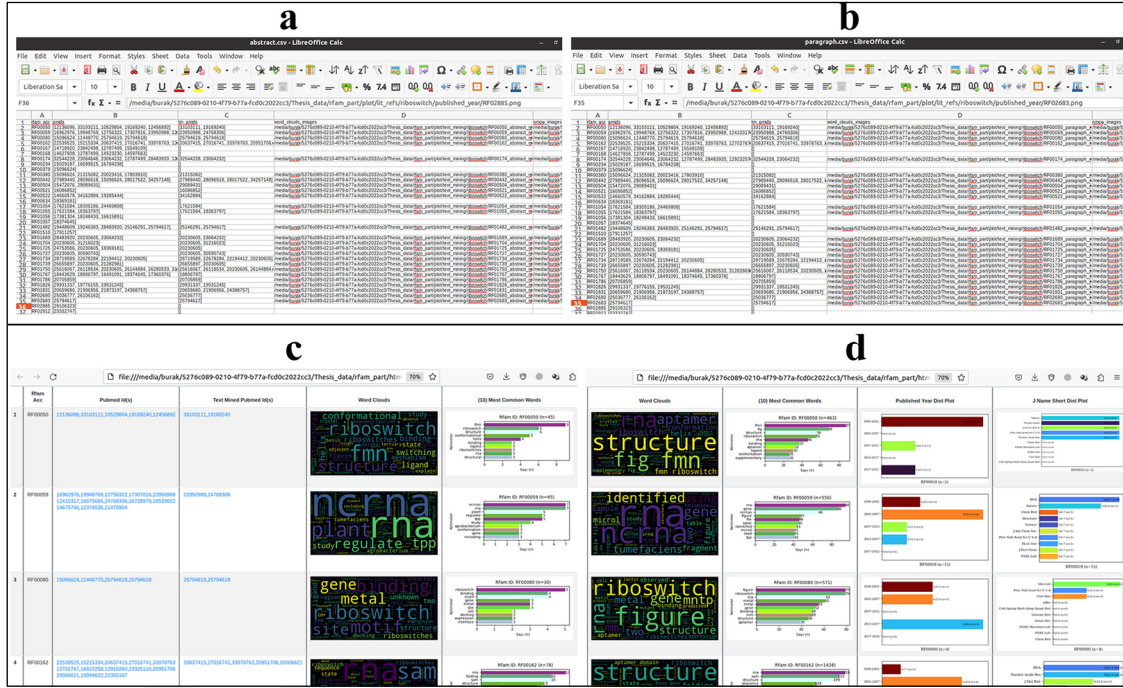
Her iki tür veri için de aynı metin madenciliği işlemleri uygulanmıştır. İlk işlem de doğal dil işleme (NLP) tekniği kullanılarak kaynak makale verileri için “kelime bulutu” (word clouds) grafikleri elde edilmiştir. Bu işlem için *Phyton* programlama dilinin “*nltk*” ve “*wordcloud*” paketlerinde yer alan çeşitli kütüphanelerdeki fonksiyonlar kullanılmıştır. İkinci işlemde ise Doğrusal Ayırma Analizi (LDA) kullanılarak kaynak makale verileri içerisinde en çok tekrar eden on kelime tespit edilmiş ve bu kelimelerin frekans dağılım grafikleri de çizdirilmiştir. Bu işlem için “*Phyton*” programlama dilinin “*Scikit-learn*” paketinde yer alan çeşitli kütüphanelerdeki fonksiyonlar kullanılmıştır (99). Bu adıma ait sürecin daha iyi anlaşılması için Resim 42 incelenebilir.



Resim 42. Rfam biyolojik veri tabanında bulunan literatür referansları veri seti üzerine metin madenciliği uygulanarak grafik çıktıların elde edilmesi sürecinde kullanılan kod betiğine ait akış sürecinin ikinci adımında yapılan işlemler. (a) Birinci analizde kullanılmak amacıyla “api” desteğiyle elde edilen ve “<pubmed referans no>.csv” formatında yerel diske kaydedilen dosyaların içeriğinden örnek bir kesit. (b) İkinci analizde kullanılmak amacıyla “api” desteğiyle elde edilen ve “<pubmed_referans_no>.csv” formatında yerel diske kaydedilen dosyaların içeriğinden örnek bir kesit. (c) Birinci analiz sonucunda elde edilen, kelime bulutu “wordcloud” grafiklerinden örnek bir kesit. (d) İkinci analiz sonucunda elde edilen, kelime bulutu “wordcloud” grafiklerinden örnek bir kesit. (e) Birinci analiz sonucunda elde edilen, en çok tekrar eden on kelime grafiklerinden örnek bir kesit. (f) İkinci analiz sonucunda elde edilen, en çok tekrar eden on kelime grafiklerinden örnek bir kesit.

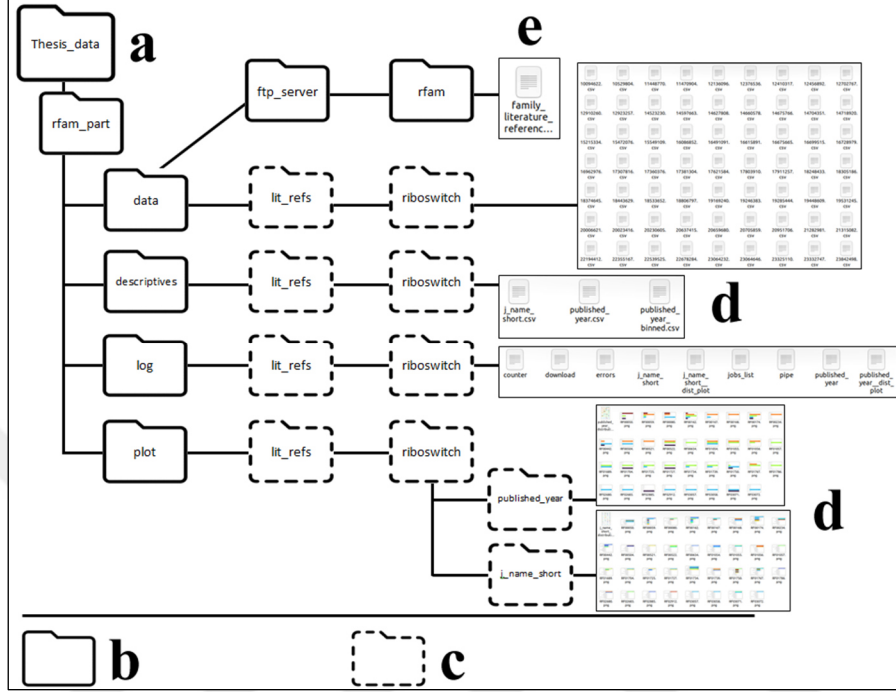
Üçüncü adım: Bu kısımda kod betiğine ait ana çıktı yani ana raporun oluşturulma işlemi gerçekleştirilmiştir. Rapor içeriğinde bir önceki başlıkta elde edilen ve kısaca “basım yılı” ve “dergi adı” olarak ifade edilen verilere ait grafik çıktılarına da yer verilmiştir. Bunlarla birlikte bu kısımda çalışan kod betiğiyle elde edilen “kelime bulutu” (word cloud) grafikleri ve en çok tekrar eden on kelime grafiklerine de yer verilmiştir.

Ayrıca araştırmacıların daha rahat inceleyebilmesi için her bir RNA elemanı için yazılmış “Pubmed” referans numaralı makalelere erişim bağlantılarına da rapor içinde yer verilmiştir. Bu adıma ait sürecin daha iyi anlaşılması için Resim 43 incelenebilir.



Resim 43. Rfam biyolojik veri tabanında bulunan literatür referansları veri seti üzerine metin madenciliği uygulanarak grafik çıktılarının elde edilmesi sürecinde kullanılan kod betiğine ait akış sürecinin üçüncü adımında yapılan işlemler. (a) Birinci analizde kullanılmak amacıyla “api” desteğiyle elde edilen ve “<pubmed_referans_no>.csv” formatında yerel diske kaydedilen dosyaların içeriğinden örnek bir kesit. (b) İkinci analizde kullanılmak amacıyla “api” desteğiyle elde edilen ve “<pubmed_referans_no>.csv” formatında yerel diske kaydedilen dosyaların içeriğinden örnek bir kesit. (c) Birinci analiz sonucunda elde edilen, kelime bulutu “wordcloud” grafiklerinden örnek bir kesit. (d) İkinci analiz sonucunda elde edilen, kelime bulutu “wordcloud” grafiklerinden örnek bir kesit. (e) Birinci analiz sonucunda elde edilen, en çok tekrar eden on kelime grafiklerinden örnek bir kesit. (f) İkinci analiz sonucunda elde edilen, en çok tekrar eden on kelime grafiklerinden örnek bir kesit.

Burada ifade edilen işlem adımlarının her biri için kullanılan dizin yapısını daha iyi anlamak için Şekil 12'yi incelemeniz önerilir.



Şekil 12. Rfam biyolojik veri tabanında bulunan literatür referansları veri seti üzerine metin madenciliği uygulanarak grafik çıktıların elde edilmesi süreci için tasarlanan dizin yapısı ve elde edilen dosyalar. (a) Kök dizini ifade eder. (b) Kod betiğinin çalışma sürecinde belirtilen kök dizin içerisinde daha önceki adımlarda oluşturulmuş klasörleri ifade eder. (c) Kod betiğinin çalışma sürecinde belirtilen kök dizin için ilk defa oluşturulan klasörleri ifade eder. (d) Kod betiği işlemini tamamladığında elde edilen çıktı, tanımlayıcı veri, grafik ve log kayıt dosyalarını ifade eder. (e) Kod betiğinin çalışması için gerekli girdi dosyasını ifade eder.

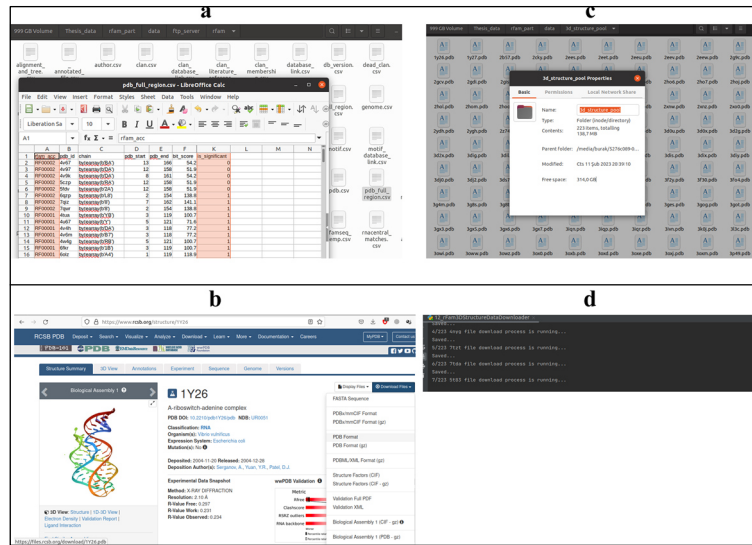
3.3.11. Rfam Biyolojik Veri Tabanında Bulunan Üç Boyutlu Yapı Veri Seti İçerisinde Yer Alan Pdb Dosyalarının İndirilmesi

Çalışmanın bu kısmıyla alakalı işlemler için yazılan kod betiğinin takip ettiği iş akışı aşağıda detaylı bir şekilde işlem adımlarına bölünerek ifade edilmeye çalışılmıştır.

Birinci adım: Daha önceki adımlarda erişime açık “Rfam” SQL sunucusu kullanılarak “Rfam” biyolojik veri tabanının güncel sürümünün dosyaları yerel diske indirilmiştir. Bu kısımda indirilmiş olan dosyalardan “pdb_full_region.csv” dosyası kod betiğinin girdi dosyası olarak kullanılmıştır. Girdi dosyasına çift katmanlı bir filtreleme işlemi uygulanmıştır. Birinci katmanda girdi dosyasının “rfam_acc” veri alanı araştırma konumuz olan RNA elemanlarına ait kayıtları kapsayacak şekilde filtrelenmiştir. İkinci

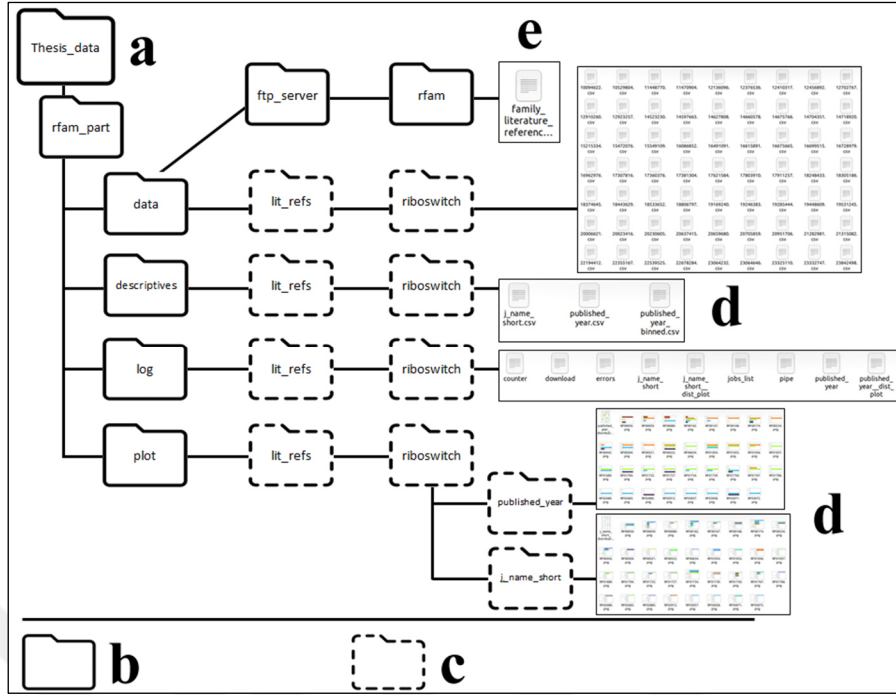
katmanda ise girdi dosyasında yer alan “*is_significant*” veri alanı kayıtlarının filtrelenmesinde kullanılmış olup sadece anlamlı kabul edilen kayıtlar bu süreçte kullanılmıştır (100). Ayrıca, filtreleme işlemlerinde kullanılan girdi dosyasında PDB tanımlayıcısı olarak ifade edebileceğimiz bir veri alanı da yer almaktadır.

PDB tanımlayıcısı, Protein Veri Tabanı içerisinde yer alan ve bu biyolojik veri tabanına özel olarak tanımlanmış, üç boyutlu yapı dosyalarına erişim için gerekli özel bir erişim kimlik verisidir (50). Açıklaması verilen bu tanımlayıcıları kullanan kod betiği Protein Veri Tabanı sunucusundan gerekli dosyaları yerel diske indirmiştir. İndirme işlemi sürecinde bazı *pdb* tanımlayıcılarına karşılık gelen dosyaların sunucuda bulunamama ihtimali de dikkate alınmıştır. İndirilen dosyaların yerel diskteki konum verileri “*download*” adlı *log* kayıt dosyasına, sunucuda bulunamayan ve dolayısıyla indirilemeyen dosyalara ait *pdb* tanımlayıcısı verileri de “*error*” adlı *log* kayıt dosyasına yazılması planlanmıştır. Bu süreçte 17 152 adet kayıttan 8 348 tanesi tekildir. Tekil olan kayıtlardan 223 adedi çift katmanlı filtreleme kuralını geçmiştir. Filtreleme kuralını geçen dosyaların tamamı yerel diske başarıyla indirilmiştir. Bu adıma ait sürecin daha iyi anlaşılması için Resim 44 incelenebilir.



Resim 44. *Rfam* biyolojik veri tabanında bulunan üç boyutlu yapı veri seti içeriğinde yer alan *pdb* dosyalarının indirilmesi sürecinde, kullanılan kod betiğine ait akış sürecinin birinci adımında yapılan işlemler. (a) Girdi dosyası olan “*pdb_full_region.csv*” yerel diskteki konumundan ve içeriğinden bir kesit. (b) Üç boyutlu yapı dosyalarının indirilebilmesi için gerekli bağlantıların bulunduğu biyolojik veri tabanı sitesinden bir kesit. (c) Yerel diske indirilen üç boyutlu yapı dosyalarının konumunu ve adedini gösteren bir kesit. (d) Kod betiğinin çalışma sürecinden bir kesit.

İkinci adım: Bu adımda yerel diske indirilmiş olan “*pdb*” dosyalar için ilave veri dosyalarının indirilmesi işlemi gerçekleştirilmiştir. Bu işlem için; Protein Veri Tabanının *ftp* sunucusundaki ilgili konumda yer alan “*idx*” uzantılı dosyalar kullanılmıştır (101). Belirtilen “*idx*” uzantılı dosyalar *ftp* sunucusundan *Python* programlama dilinin “*ftplib*” paketinde yer alan “*FTP*” kütüphanesindeki çeşitli fonksiyonlar kullanılarak yerel diske indirilmiştir. İndirilmiş olan “*idx*” uzantılı dosyalara ait veri alanları, kod betiğinde tanımlanan yönteme göre okunup derlenerek “*csv*” uzantılı iki boyutlu tablo düzenine çevrilmiştir. Kısa açıklaması verilen bu yöntem doğrultusunda gerekli ilave veri dosyalarına ulaşılmıştır. Bu süreçte indirilen “*idx*” uzantılı dosyalardan “*cmpd_res.idx*”, “*entries.idx*” ve “*source.idx*” adlı dosyalar sonraki işlemlerde ilave veri olarak kullanılabilmek amacıyla “*csv*” uzantılı “*cmpd_res.csv*”, “*entries.csv*” ve “*source.csv*” sürümlerine dönüştürülmüştür. Bu adıma ait sürecin daha iyi anlaşılması için Resim 45 incelenebilir.



Şekil 13. Rfam biyolojik veri tabanında bulunan üç boyutlu yapı veri seti içeriğinde yer alan *pdb* dosyalarının indirilmesi süreci için tasarlanan dizin yapısı ve elde edilen dosyalar. (a) Kök dizini ifade eder. (b) Kod betiğinin çalışma sürecinde belirtilen kök dizin içerisinde daha önceki adımlarda oluşturulmuş klasörleri ifade eder. (c) Kod betiğinin çalışma sürecinde belirtilen kök dizin için ilk defa oluşturulan klasörleri ifade eder. (d) Kod betiği işlemini tamamladığında elde edilen çıktı, tanımlayıcı veri, grafik ve *log* kayıt dosyalarını ifade eder.

3.3.12. Rcsb Biyolojik Veri Tabanından Elde Edilen İlave Veri Dosyalarıyla Veri Zenginleştirilmesi Yapılması

Çalışmanın bu kısmıyla alakalı işlemler için yazılan kod betiğinin takip ettiği iş akışı aşağıda detaylı bir şekilde işlem adımlarına bölünerek ifade edilmeye çalışılmıştır.

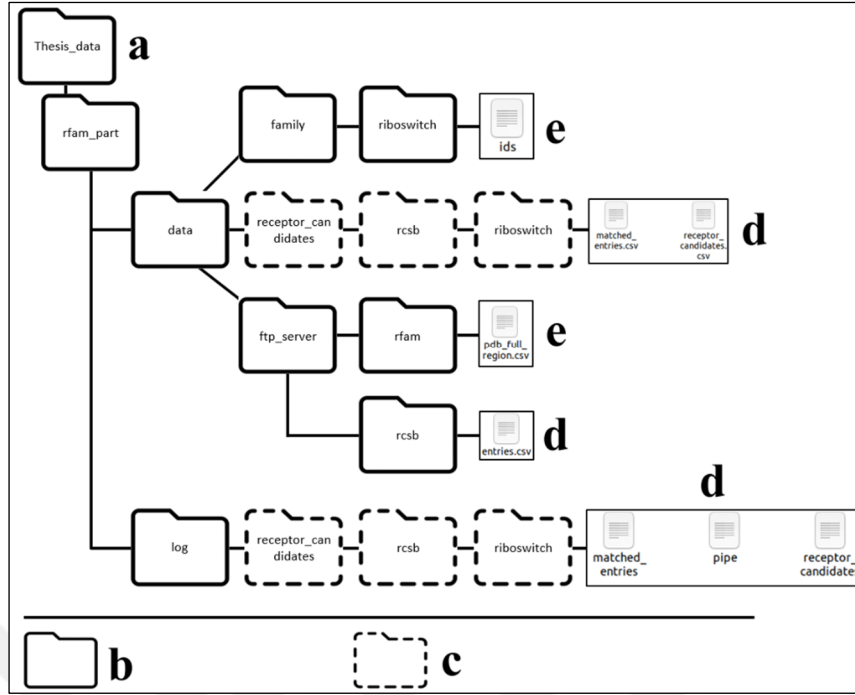
Birinci adım: Bu adımda, bir önceki adımda elde edilen zenginleştirme dosyalarından, “*entries.csv*” dosyasında yer alan kayıtlar kullanılmıştır. Bir önceki başlıkta filtrelene yöntemi anlatılan “*pdb_full_region.csv*” dosyasındaki benzersiz (tekil) *pdb* tanımlayıcı verileri ile “*entries.csv*” dosyasındaki kayıtlar eşleştirilerek “*matched_entries.csv*” dosyası elde edilmiştir. Böylelikle “*pdb_full_region.csv*” içerisindeki kısıtlı veri yerine, zenginleştirilmiş veri içeriğine sahip “*matched_entries.csv*” dosya içeriği kod betiğinin sonraki adımları için kullanılacaktır.

İkinci adım: Bu adımda, bir önceki adımda oluşturulan “*matched_entries.csv*” dosyası ve birinci adımda indirilen “*pdb*” dosyalarının yerel diskteki konum bilgileri girdi

verisi olarak kullanılmıştır. Bahsi geçen “*csv*” dosyası içerisinde yer alan “*SOURCE*” veri alanı her bir kayıta yer alan “*pdb*” dosya tanımlayıcısının kaynak genomuna ait veriyi içermektedir. Ancak bu veri, canlıların bilimsel sınıflandırmasında kullanılan taksonomi terimi karşılığını içermesine rağmen gereksiz bazı ilave verileri de içerisinde barındırmaktadır. Bu nedenle burada yapılan ilk işlem, verideki gereksiz kısımların temizlenmesidir. Kod betiği bu amaçla veri içerisindeki örüntüyü keşfederek ve bazı metin fonksiyonlarını da kullanarak gerekli temizleme işlemini yerine getirmiştir.

Genom adlarının bilimsel karşılıkları başarıyla elde edildikten sonra; “*matched_entries.csv*” dosyasındaki “*COMPOUND*” veri alanında yer alan kayıtlarda sadece araştırma konumuz olan Riboşalter kelimesinin geçtiği kayıtların saklandığı bir filtreleme işlemi daha uygulanmıştır. Böylelikle yerel diske indirilmiş olan 223 adet üç boyutlu yapı dosyası içerisinde, Riboşalter *anotasyon* verisi olan 38 adet dosya üzerinden çalışmaya devam edilmesi kararlaştırılmıştır. Ardından zenginleştirilmiş ve araştırma konumuza göre filtrelenmiş kayıtlardan oluşan dosya başarıyla oluşturulmuştur. Oluşturulan bu dosya “*receptor_candidates.csv*” adıyla yerel diske yazılmıştır.

İfade etmeye çalıştığımız süreç hakkında daha iyi fikir sahibi olabilmek için Resim 46 ve Şekil 14 incelemeniz önerilir.



Şekil 14. *Rcsb* biyolojik veri tabanından elde edilen ilave veri dosyalarıyla veri zenginleştirilmesi yapılması süreci için tasarlanan dizin yapısı ve elde edilen dosyalar. (a) Kök dizini ifade eder. (b) Kod betiğinin çalışma sürecinde belirtilen kök dizin içerisinde daha önceki adımlarda oluşturulmuş klasörleri ifade eder. (c) Kod betiğinin çalışma sürecinde belirtilen kök dizin için ilk defa oluşturulan klasörleri ifade eder. (d) Kod betiği işlemini tamamladığında elde edilen çıktı ve *log* kayıt dosyalarını ifade eder. (e) Kod betiğinin çalışması için gerekli girdi dosyasını ifade eder.

3.3.13. Reseptör Adayı Üç Boyutlu Yapı Kayıtlarını Tutan Dosyanın Sonraki Süreçler İçin Raporlanması

Çalışmanın bu kısmıyla alakalı işlemler için yazılan kod betiğinin takip ettiği iş akışı aşağıda detaylı bir şekilde işlem adımlarına bölünerek ifade edilmeye çalışılmıştır.

Birinci adım: Bu adımda, bir önceki adımda elde edilen “*receptor_candidates.csv*” adlı çıktı dosyasına ait kayıtlar kod betiği için girdi olarak kullanılmıştır. Girdi kayıtlarındaki “*organism_name*” veri alanında yer alan ve kaynak genom adlarından oluşan her bir kayıt kullanılarak ilave veri zenginleştirilmesi yapılmıştır. Bu işlem için “*Phyton*” programlama dilinin “*BioPhyton*” (92) paketi altındaki “*Entrez*” kütüphanesindeki çeşitli fonksiyonlar kullanılmıştır. Bu fonksiyonlar yardımıyla klasör adlarında yer alan kaynak genomların taksonomi numaraları ve taksonomi soy ağacı (*lineage*) verileri elde edilmiştir. Kod betiği sonucunda elde edilen çıktı dosyası yerel diske “*entrez_receptor_candidates.csv*” adıyla kaydedilmiştir. Bu adıma ait sürecin daha iyi anlaşılması için Resim 47 incelenebilir.

a				b			
receptor_candidates.csv				entrez_receptor_candidates.csv			
rfam_acc	id_code	organism_name	annotation	id_code	organism_name	tax_id	message
RF00379	6NSQ	homo sapiens	Structure of Human pi-c-mRNA-378a Apical Loop and One	1	id_code	9606	cellular organisms; Eukaryota; Opisthokonta; Metazoa; Eum
RF00167	1Y27	bacillus subtilis	O-ribosyltransferase complex	2	6NSQ	9606	cellular organisms; Eukaryota; Opisthokonta; Metazoa; Eum
RF00168	4ERJ	thermotoga maritima	Crystal structure of the lysine riboswitch bound to a 6-am	3	1Y27	1423	cellular organisms; Bacteria; Terrabacteria group; Bacilota;
RF00059	3D2X	arabidopsis thaliana	Structure of the thiamine pyrophosphate-specific ribosw	4	4ERJ	2336	cellular organisms; Bacteria; Thermotoga; Thermotoga; T
RF00379	4W90	bacillus subtilis	Crystal structure of Bacillus subtilis cyclic-di-AMP ribosw	5	3D2X	3702	cellular organisms; Eukaryota; Viridiplantae; Streptophyta; S
RF02680	4RZD	homo sapiens	Crystal structure of a PheQ1 Riboswitch	6	4W90	1423	cellular organisms; Bacteria; Terrabacteria group; Bacilota;
RF00168	4ERL	thermotoga maritima	Crystal structure of the lysine riboswitch bound to a lysine	7	4RZD	9606	cellular organisms; Eukaryota; Opisthokonta; Metazoa; Eum
RF00379	6NS5	homo sapiens	Structure of Human pi-c-mRNA-320b-2 Apical Loop and O	8	4ERL	2336	cellular organisms; Bacteria; Thermotoga; Thermotoga; T
RF00379	6NSO	homo sapiens	Structure of Human pi-c-mRNA-202 Apical Loop and One	9	6NS5	9606	cellular organisms; Eukaryota; Opisthokonta; Metazoa; Eum
RF00504	3P49	faecalibacterium nucleatum	Crystal Structure of a Glycine Riboswitch from Faecalibact	10	6NSO	9606	cellular organisms; Eukaryota; Opisthokonta; Metazoa; Eum
RF00504	3P49	homo sapiens	Crystal Structure of a Glycine Riboswitch from Faecalibact	11	3P49	851	cellular organisms; Bacteria; Fusobacteri; Fusobacteri; F
RF00379	4QLM	thermoanaerobacter tengcongensis	ytidao riboswitch binding to c-di-AMP	12	4QLM	9606	cellular organisms; Eukaryota; Opisthokonta; Metazoa; Eum
RF00059	3D2G	arabidopsis thaliana	Structural basis of thiamine pyrophosphate analogues bin	13	3D2G	119072	cellular organisms; Bacteria; Terrabacteria group; Bacilota;
RF00167	1Y26	vibrio vulnificus	A-riboswitch-adenosine complex	14	3D2G	3702	cellular organisms; Eukaryota; Viridiplantae; Streptophyta; S
RF00059	3D2V	arabidopsis thaliana	Structure of the thiamine pyrophosphate-specific riboswitch bound	15	1Y26	672	cellular organisms; Bacteria; Pseudomonadota; Gammaproteobact
RF00379	6NSN	homo sapiens	Structure of Human pi-c-mRNA-208a Apical Loop and One	16	3D2V	3702	cellular organisms; Eukaryota; Viridiplantae; Streptophyta; S
RF00379	6NSP	homo sapiens	Structure of Human pi-c-mRNA-340 Apical Loop and One	17	6NSN	9606	cellular organisms; Eukaryota; Opisthokonta; Metazoa; Eum
RF00379	6WTL	homo sapiens	Structure of Human pi-c-mRNA-190-2 Apical Loop and On	18	6NSP	9606	cellular organisms; Eukaryota; Opisthokonta; Metazoa; Eum
RF00379	4W92	bacillus subtilis	Crystal structure of Bacillus subtilis cyclic-di-AMP ribosw	19	6WTL	9606	cellular organisms; Eukaryota; Opisthokonta; Metazoa; Eum
RF00379	4W92	homo sapiens	Crystal structure of Bacillus subtilis cyclic-di-AMP ribosw	20	4W92	1423	cellular organisms; Bacteria; Terrabacteria group; Bacilota;
RF00379	6WTR	homo sapiens	Structure of Human pi-c-mRNA-300 Apical Loop Fused to	21	4W92	9606	cellular organisms; Eukaryota; Opisthokonta; Metazoa; Eum
RF00167	3IVN	bacillus subtilis	Structure of the U55C mutant A-riboswitch aptamer from	22	6WTR	9606	cellular organisms; Eukaryota; Opisthokonta; Metazoa; Eum
RF00379	6NST	homo sapiens	Structure of Human pi-c-mRNA-378a Apical Loop Fused ti	23	3IVN	1423	cellular organisms; Bacteria; Terrabacteria group; Bacilota;
RF01831	6Q57	streptococcus rogersi	X-ray crystal structure of the tetrahydrofolate riboswitch s	24	6NST	9606	cellular organisms; Eukaryota; Opisthokonta; Metazoa; Eum
				25	6Q57	1309	cellular organisms; Bacteria; Terrabacteria group; Bacilota;
				26	6Q57	9606	cellular organisms; Eukaryota; Opisthokonta; Metazoa; Eum
				27	6Q57	1224	cellular organisms; Bacteria
				28	6HAG		

c

```

14_rfam3DStructureDataDescriptiver x
3702 <----> arabidopsis thaliana
5/33 step is running...
1423 <----> bacillus subtilis
6/33 step is running...
9606 <----> homo sapiens
7/33 step is running...
853 <----> faecalibacterium prausnitzii

```

Resim 47. Reseptör Adayı Üç Boyutlu Yapı Kayıtlarını Tutan Dosyanın Sonraki Süreçler İçin Raporlanması sürecinde, kullanılan kod betiğine ait akış sürecinin birinci adımında yapılan işlemler. (a) Kod betiği girdisi olarak kullanılan “receptor_candidates.csv” adlı dosya içeriğini ve ilave veri zenginleştirilmesi işleminde kullanılan veri alanını gösteren bir kesit. (b) İlave veri zenginleştirilmesi yapılan içeriği barındıran “entrez_receptor_candidates.csv” adlı dosya içeriğinden bir kesit. (c) Kod betiği çalışma sürecinden bir kesit.

İkinci adım: Bu adımda kod betiği bir önceki adımda elde edilen “entrez_receptor_candidates.csv” dosya içeriğinde yer alan ve standart bir düzende olmayan taksonomi soy ağacı (lineage) verileri içerisinde keşfedilen örüntüyü kullanarak her bir kayıt için “superkingdom” taksonomi düzeyine karşılık gelen veriyi basit metin işlevleriyle çeker ve çektiği bu veriyi “superkingdom” adlı veri alanına yazar. Bu yeni veri alanına sahip dosya, “superkingdom_level_enriched.csv” adıyla yerel diske kaydedilmiştir.

Ardından yerel diske kaydedilmiş olan bu dosyadaki “superkingdom” adlı veri alanında yer alan kayıtların araştırma konumuz olan RNA elemanlarının Rfam biyolojik veri tabanı erişim kimliklerine göre dağılımı hesaplanmıştır. Dağılım verilerini tutan bu dosya da yerel diskte ilgili konumuna “descriptive.csv” adıyla kaydedilmiştir. Bu adıma ait sürecin daha iyi anlaşılması için Resim 48 incelenebilir.

a

Accession	Organism	Description
1	3702	cellular organisms; Eukaryota; Viridiplantae; Streptophyta; S
2	3702	cellular organisms; Eukaryota; Viridiplantae; Streptophyta; S
3	3702	cellular organisms; Eukaryota; Viridiplantae; Streptophyta; S
4	3702	cellular organisms; Eukaryota; Viridiplantae; Streptophyta; S
5	3702	cellular organisms; Eukaryota; Viridiplantae; Streptophyta; S
6	3702	cellular organisms; Eukaryota; Viridiplantae; Streptophyta; S
7	3702	cellular organisms; Eukaryota; Viridiplantae; Streptophyta; S
8	3702	cellular organisms; Eukaryota; Viridiplantae; Streptophyta; S
9	3702	cellular organisms; Eukaryota; Viridiplantae; Streptophyta; S
10	3702	cellular organisms; Eukaryota; Viridiplantae; Streptophyta; S
11	3702	cellular organisms; Eukaryota; Viridiplantae; Streptophyta; S
12	3702	cellular organisms; Eukaryota; Viridiplantae; Streptophyta; S
13	3702	cellular organisms; Eukaryota; Viridiplantae; Streptophyta; S
14	3702	cellular organisms; Eukaryota; Viridiplantae; Streptophyta; S
15	3702	cellular organisms; Eukaryota; Viridiplantae; Streptophyta; S
16	3702	cellular organisms; Eukaryota; Viridiplantae; Streptophyta; S
17	3702	cellular organisms; Eukaryota; Viridiplantae; Streptophyta; S
18	3702	cellular organisms; Eukaryota; Viridiplantae; Streptophyta; S
19	3702	cellular organisms; Eukaryota; Viridiplantae; Streptophyta; S
20	3702	cellular organisms; Eukaryota; Viridiplantae; Streptophyta; S
21	3702	cellular organisms; Eukaryota; Viridiplantae; Streptophyta; S
22	3702	cellular organisms; Eukaryota; Viridiplantae; Streptophyta; S
23	3702	cellular organisms; Eukaryota; Viridiplantae; Streptophyta; S
24	3702	cellular organisms; Eukaryota; Viridiplantae; Streptophyta; S
25	3702	cellular organisms; Eukaryota; Viridiplantae; Streptophyta; S
26	3702	cellular organisms; Eukaryota; Viridiplantae; Streptophyta; S
27	3702	cellular organisms; Eukaryota; Viridiplantae; Streptophyta; S
28	3702	cellular organisms; Eukaryota; Viridiplantae; Streptophyta; S
29	3702	cellular organisms; Eukaryota; Viridiplantae; Streptophyta; S
30	3702	cellular organisms; Eukaryota; Viridiplantae; Streptophyta; S
31	3702	cellular organisms; Eukaryota; Viridiplantae; Streptophyta; S
32	3702	cellular organisms; Eukaryota; Viridiplantae; Streptophyta; S
33	3702	cellular organisms; Eukaryota; Viridiplantae; Streptophyta; S
34	3702	cellular organisms; Eukaryota; Viridiplantae; Streptophyta; S

b

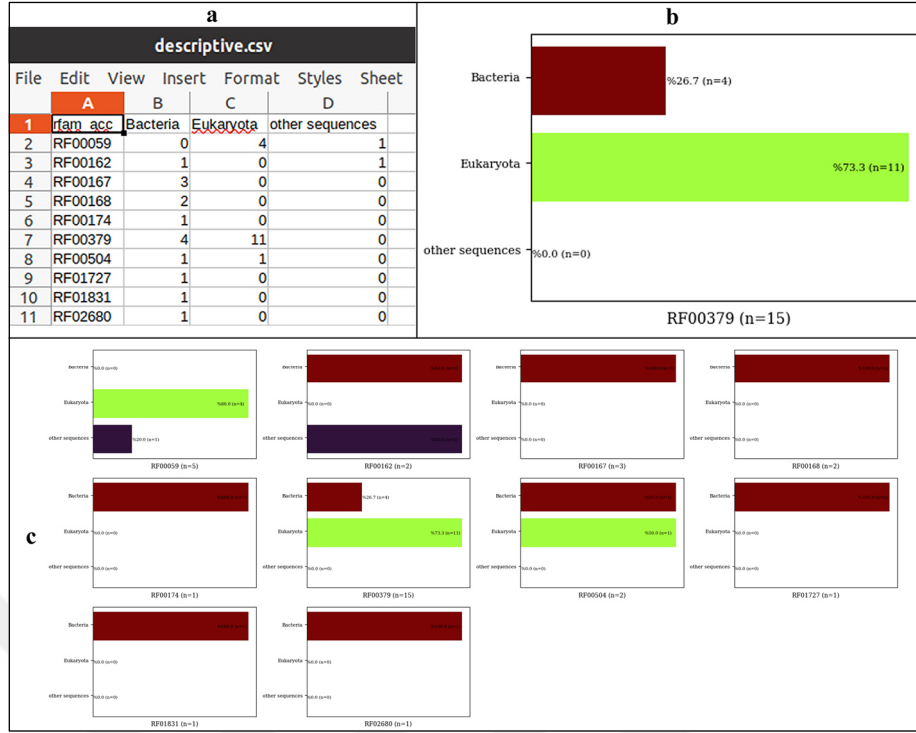
Accession	Organism	Superkingdom
1	3702	Eukaryota
2	3702	Eukaryota
3	3702	Eukaryota
4	3702	Eukaryota
5	3702	Eukaryota
6	3702	Eukaryota
7	3702	Eukaryota
8	3702	Eukaryota
9	3702	Eukaryota
10	3702	Eukaryota
11	3702	Eukaryota
12	3702	Eukaryota
13	3702	Eukaryota
14	3702	Eukaryota
15	3702	Eukaryota
16	3702	Eukaryota
17	3702	Eukaryota
18	3702	Eukaryota
19	3702	Eukaryota
20	3702	Eukaryota
21	3702	Eukaryota
22	3702	Eukaryota
23	3702	Eukaryota
24	3702	Eukaryota
25	3702	Eukaryota
26	3702	Eukaryota
27	3702	Eukaryota
28	3702	Eukaryota
29	3702	Eukaryota
30	3702	Eukaryota
31	3702	Eukaryota
32	3702	Eukaryota
33	3702	Eukaryota
34	3702	Eukaryota

c

Accession	Bacteria	Eukaryota	other sequences
1	0	4	1
2	0	1	1
3	0	0	0
4	0	0	0
5	0	0	0
6	0	0	0
7	0	11	0
8	0	1	0
9	0	0	0
10	0	0	0
11	0	0	0

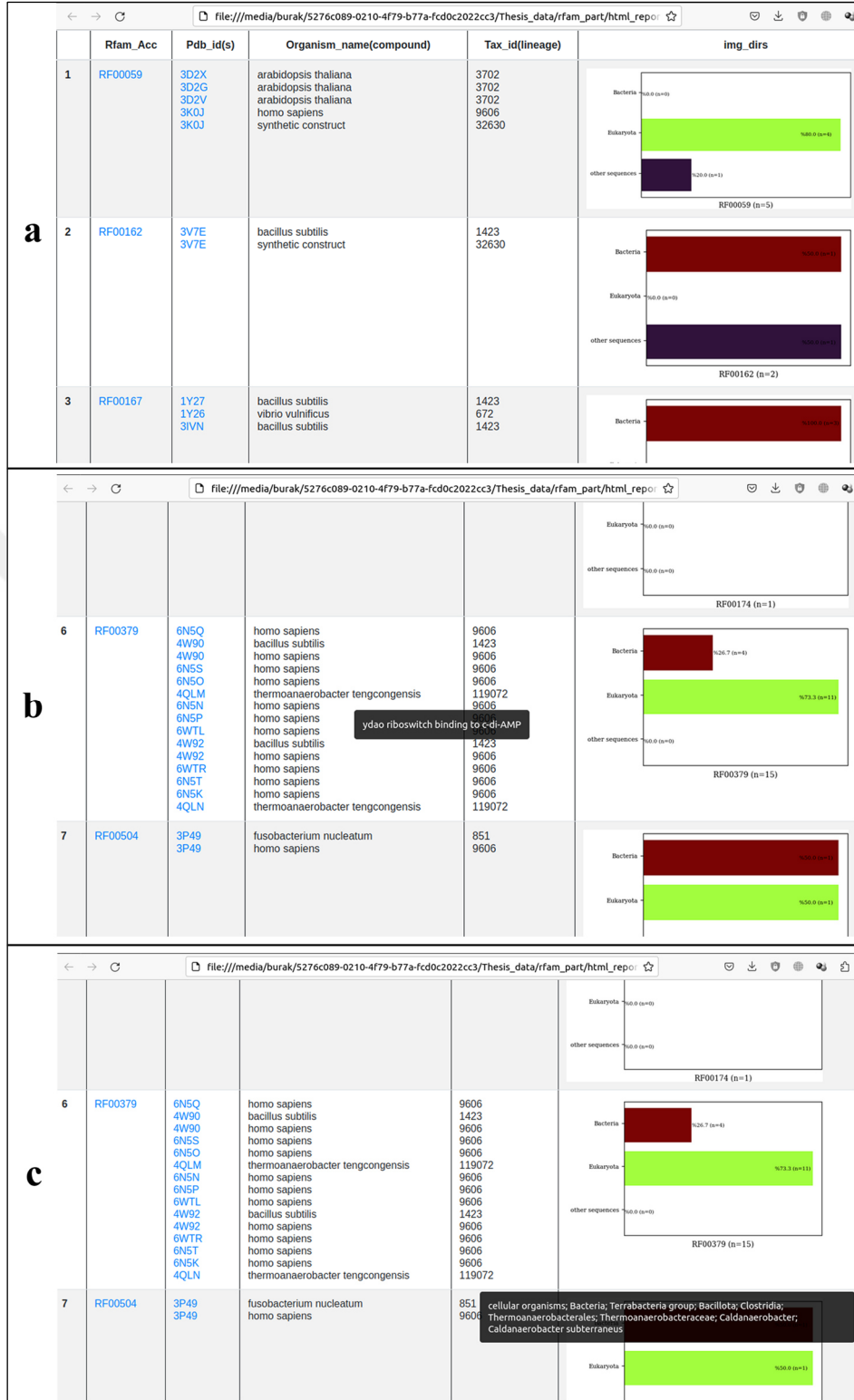
Resim 48. Reseptör Adayı Üç Boyutlu Yapı Kayıtlarını Tutan Dosyanın Sonraki Süreçler İçin Raporlanması sürecinde, kullanılan kod betiğine ait akış sürecinin ikinci adımında yapılan işlemler. **(a)** Grafik çıktısı elde etmek için örüntü keşfi yapılan “entrez_receptor_candidates.csv” adlı dosyadaki veri alanını gösteren bir kesit. **(b)** Grafik çıktısı elde etmek için oluşturulan ve “superkingdom” veri alanına sahip olan “superkingdom_level_enriched.csv” adlı dosya içeriğinden bir kesit. **(c)** Grafik çıktısı elde etmek için gerekli frekans dağılım verilerini içerecek şekilde oluşturulan “descriptive.csv” adlı dosya içeriğinden bir kesit.

Üçüncü adım: Bu adımda, girdi olarak bir önceki adımlarda oluşturulan “descriptive.csv” dosyası kullanılmıştır. Bu dosya kullanılarak; reseptör adayları olarak belirlenen üç boyutlu yapı dosyalarının kaynağı olan genomların “superkingdom” düzeyindeki taksonomi verisine göre frekans dağılım grafikleri yatay çubuk grafiği türünde oluşturulmuştur. Son olarak ise oluşturulan tüm grafikler toplu halde tek bir grafik çıktısı olarak elde edilmiştir. Bu adımda elde edilen çıktıların Resim 49 içerisinde yer verilmiştir.



Resim 49. Reseptör Adayı Üç Boyutlu Yapı Kayıtlarını Tutan Dosyanın Sonraki Süreçler İçin Raporlanması sürecinde, kullanılan kod betiğine ait akış sürecinin üçüncü adımında yapılan işlemler. **(a)** Grafik çıktısı elde etmek için kullanılan şekilde oluşturulan “*descriptive.csv*” adlı dosya içeriğinden bir kesit. **(b)** Tek bir RNA elemanına ait veriler içeren düzenlenmiş frekans dağılım grafiğinden örnek bir kesit. **(c)** Araştırma kapsamına giren tüm RNA elemanlarına ait veriler içeren düzenlenmiş frekans dağılım grafiklerinin toplu gösteriminden bir kesit.

Dördüncü adım: Kod betiğinin son bölümünü oluşturan bu kısımda reseptör adayları olarak belirlenen üç boyutlu yapılar raporlanmıştır. Ayrıca araştırmacıların reseptör adayları üç boyutlu yapıları daha iyi inceleyebilmesi için gerekli erişim bağlantıları da verilmiştir (50). Rapor dosyası içerisinde daha düzenli bir görünüm elde edebilmek için bazı veriler direkt raporda verilmemiş; bunun yerine etkileşimli içerik olarak ayarlanmışlardır. Örneğin; araştırmacı, taksonomi kimliği kayıtlarının üzerine geldiğinde taksonomi soyağacı verisini de görebilmektedir. Araştırma kapsamına giren her bir RNA elemanına ait yatay çubuk grafiği de rapor dosyası içerisinde konumlandırılmıştır. İçeriğinden bahsettiğimiz sonuç rapor dosyası “*html*” formatında dosya olarak diske yazılmıştır. Bu adıma ait sürecin daha iyi anlaşılması için Resim 50 incelenebilir.



Resim 50. Reseptör adayı üç boyutlu yapı kayıtlarını tutan dosyanın sonraki süreçler için raporlanması sürecinde, kullanılan kod betiğine ait akış sürecinin dördüncü adımında yapılan işlemler. **(a)** Kod betiği iş akışı sonucunda elde edilen “html” rapor dosyasından bir kesit. **(b)** Elde edilen “html” rapor içerisindeki “organism_name” veri alanındaki kayıtların etkileşimli içeriğine dair örnek gösterimden bir kesit. **(c)** Elde edilen “html” rapor içerisindeki “tax_id” veri alanındaki kayıtların etkileşimli içeriğine dair örnek gösterimden bir kesit.

(aşağıya bakın), ancak nihai toplamın 40'ı geçmesini beklemiyoruz. Her kaynak, yardım uç noktası ile çevrimiçi olarak belgelenmiştir (103).”

Kaynak olarak bahsedilen içeriklerin ne olduğunu daha iyi anlamak için Şekil 16'ya bakılabilir.

Resource	Is a set of	Approximate Number	Example use of this resource when the answer to your question is: a) a list of these things b) one of these things, or c) things derived from a single one of these things.
substances	molecules	200,000,000	a) compounds you can buy [1] b) ZINC ID 53 [2] c) All genes hit by compound ZINC597013 [3]
catalogs	vendor and annotated catalogs	400	a) catalogs whose members are biogenic (but see also endogenous, metabolite) [4] b) The ChEMBL20 catalog [5] c) Items in the DrugBank FDA catalogs [6] Interestingly, this shows clearly when a DrugBank ID has been expanded into multiple ZINC IDs.
genes	UniProt Gene Symbols	2800	a) All Class A GPCR genes [7] b) The BTK gene [8] c) SEA Predicted ligands for 5-HT2A [9]
catitems	What vendors and annotated catalogs call the molecules in their source catalogs	1 billion	a) All molecules where the original supplier code begins with B1234 [10] b) N.B. There is no unique key for supplier codes, so you cannot necessarily select one. The purpose of the catitem code allows you to write queries that originate in another catalog. Items in the HMDB database, where the catalog code starts with HMDB1 and the gene it is active against starts with ADR. [11] Compounds sold by Frontier, beginning with B123 and containing a pyridine ring. [12]
protomers	3D representations	6 million and growing	a) Molecules with a charge of minus 2 at pH 7.4 [13] b) One individual representation in 3D SDF format [14] c) Catalogs that this molecule can be found in [15]
orthologs	UniProt accession codes, thus species specific	3800	a) Information about orthologs by uniprot code in ZINC as a text file [16] b) Information about the FTO_HUMAN ortholog [17] c) Substances annotated to bind FTO_HUMAN [18]
activities	best ligand-gene affinity. See observations for each individual value	220,000	a) Individual data points of compound-gene associations [19] c) All activities better than 1 nM [20]
majorclasses	Major classes	15	a) A list of major classes [21] b) The membrane_receptor major class [22] c) genes in the enzyme major class [23] Another examples: substances in the genes in the major class of ion channels [24]
subclasses	Subclasses	44	a) A list of subclasses [25] b) The GPCR-A subclass [26] c) The substances that bind the GPCR F subclass [27]
organisms	asdf asdf	5	a) List of supported organisms [28] b) not supported c) not supported

Şekil 16. <https://wiki.docking.org/index.php/Resources> adresinde yer alan kaynak tablosu

Şekil 16'da içeriği verilen kaynak tablosunda yer alan içerik alt kümelerinden “protomers” alt kümesi çalışmamızda kullanılmak üzere seçilmiştir. Çünkü bu türde yer alan içerikler sanal tarama analizi için gerekli ilaç adayı moleküllerin üç boyutlu (docking işlemi için hazır) dosyalarını barındırmaktadır.

ZINC veri tabanı içerisinde yer alan veriler özel bir arşivleme mantığıyla gruplandırılmıştır. Bu gruplandırmanın yapılış mantığı özetle; “fiziksel özellik” olarak ifade edebileceğimiz özellik uzayında her bir ilaç adayı molekülün kendi alt grubunda temsil edilmesi olarak düşünülebilir. Detaylı bilgi almak isteyen araştırmacılar (104) referans numaralı makaleyi inceleyebilir.

Tablo 7. ZINC veri tabanında arşivleme için kullanılan fiziksel özellik uzayı

		A	B	C	D	E	F	G	H	I	J	K
		<-1	-1 - 0	0 - 1	1 - 2	2 - 2,5	2,5 - 3	3 - 3,5	3,5 - 4	4 - 4,5	4,5 - 5	> 5
A	< 200	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK
B	200 - 250	BA	BB	BC	BD	BE	BF	BG	BH	BI	BJ	BK
C	250 - 300	CA	CB	CC	CD	CE	CF	CG	CH	CI	CJ	CK
D	300 - 325	DA	DB	DC	DD	DE	DF	DG	DH	DI	DJ	DK
E	325 ile 350	EA	EB	EC	ED	EE	EF	EG	EH	EI	EJ	EK
F	350 ile 375	FA	FB	FC	FD	FE	FF	FG	FH	FI	FJ	FK
G	375 ile 400	GA	GB	GC	GD	GE	GF	GG	GH	GI	GJ	GK
H	400 ile 425	HA	HB	HC	HD	HE	HF	HG	HH	HI	HJ	HK
I	425 ile 450	IA	IB	IC	ID	IE	IF	IG	IH	II	IJ	IK
J	450 ile 500	JA	JB	JC	JD	JE	JF	JG	JH	JI	JJ	JK
K	> 500	KA	KB	KC	KD	KE	KF	KG	KH	KI	KJ	KK

Tablo 7’de yer alan veri diziliminde; birinci boyut olan satırlardaki değerler molekül ağırlığını, ikinci boyut olan sütunlar ise “*xlogP*” değerlerini ifade eder. Detaylı bilgi için <https://zinc15.docking.org/tranches/home/> adresi incelenebilir. Bu bilgiler doğrultusunda ve (104) referans numaralı makalede yer alan “*Obtaining 3D Models*” başlığında yer alan işlem adımlarının benzerini takip eden kod betikleri “*Phyton*” programlama dili kullanılarak yazar tarafından oluşturulmuştur.

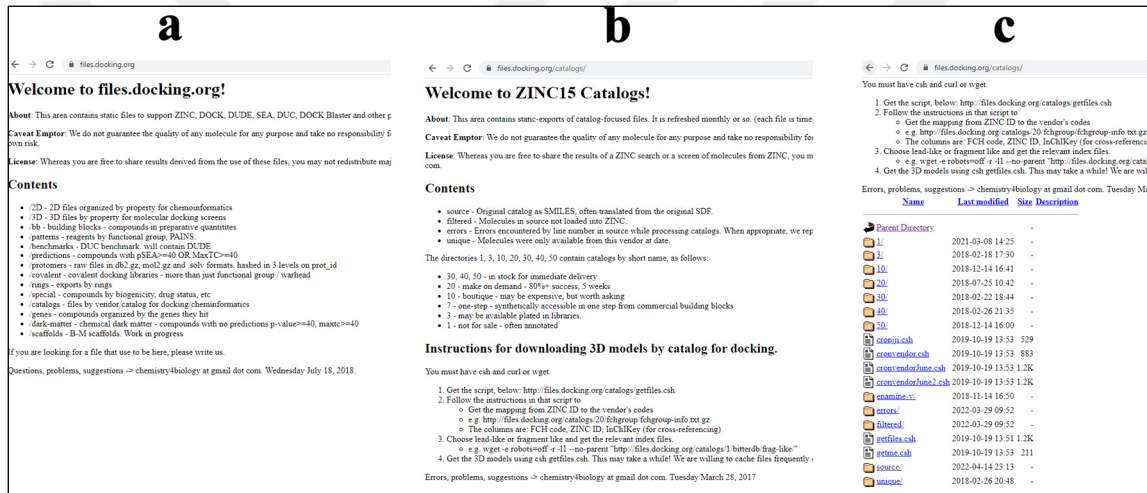
Bu bölüme ait gerekli dosyaların elde edilebilmesi için yazar tarafından oluşturulan kod betikleri ve yaptıkları işlemlere ait kısa açıklamalar tabloda verilmiştir. Tablo 8’ de yer alan her bir kod betiğinin takip ettiği iş akış sürecinde yapılan işlemler ise bu bölüme ait takip eden alt bölüm başlıkları halinde detaylı olarak ifade edilmeye çalışılmıştır.

Tablo 8. Rfam veri tabanı kısmı için oluşturulan kod betikleri

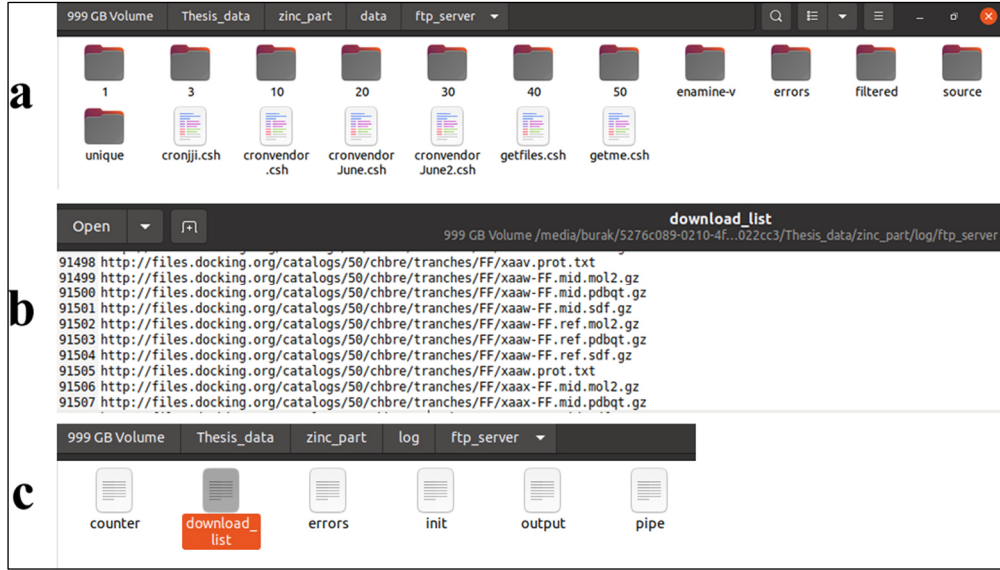
Kod betiğinin adı	Yaptığı işlem
0_dbDownloader.py	Kaynak veri tabanı sunucu dizini dosyalarını yerel diske indirme
1_dbBuilder.py	İndirilen dosyalardan gerekli kısımların alınarak ligand adayı molekül havuzu veri tabanının ön verisinin oluşturulması
2_ligandsBuilder.py	Sıkıştırılmış arşiv dosyası formatında elde edilen ilaç adayı molekül dosyalarının tek molekül dosyasına çevrilmesi ve sanal tarama analizi girdisi olabilmeleri için gerekli format dönüşümlerinin yapılması
3_dbVisualiserAndDescriptiver.py	Her bir ZINC kataloğuna ait fiziksel özellik uzayı verilerinin frekans dağılım değerlerinin sayılması ve grafik çıktısı olarak düzenlenmesi

3.4.1. ZINC Biyolojik Veri Tabanı Statik Dosyalarının Yerel Diske İndirilmesi

Bu adımda ZINC veri tabanında yer alan dosyalar, klasörler ile bu klasörlere ait alt klasör ve bunların içerisinde yer alan dosyaların yerel diske birebir senkronize olacak şekilde indirilebilmesi için gerekli log kayıt dosyasının oluşturulma işlemi gerçekleştirilmiştir. Elde edilen log kayıt dosyası “*download_list*” adıyla yerel diskte oluşturulmuştur. Ardından indirilecek dosyaların listesini tutan “*download_list*” adlı girdi dosyası kullanılarak ilgili dosyaların indirilmesi ve mevcut veri tabanının yerel diskte bir kopyasının elde edilmesi işlemleri gerçekleştirilmiştir. Bu adıma ait sürecin daha iyi anlaşılması için Resim 51 ve 52 incelenebilir.

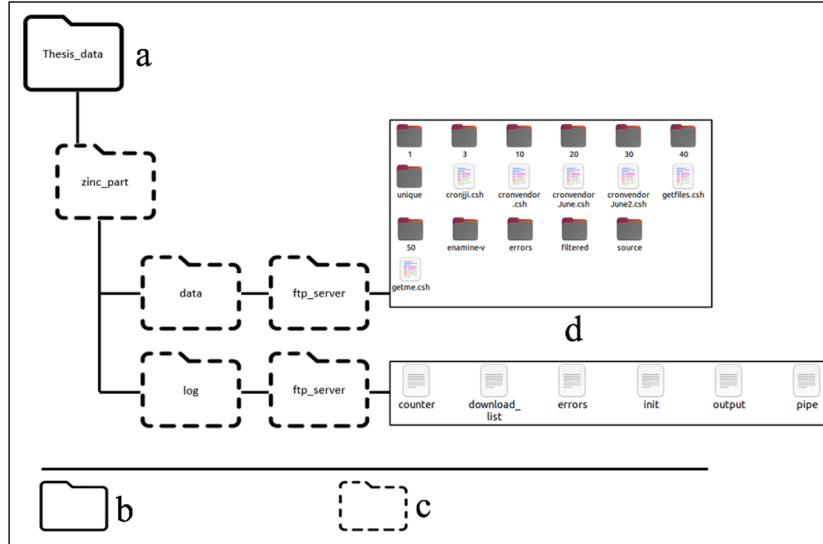


Resim 51. İlaç adayı moleküllerin indirilebilmesi için gerekli metin girdi dosya içeriğinin oluşturulma süreci. (a) ZINC biyolojik veri tabanının statik dosyalarını barındıran <https://files.docking.org/> alanına ait anasayfa içeriğinden bir kesit. (b) <https://files.docking.org/catalogs/> alanına ait sayfa içeriğinin üst kısmından bir kesit. (c) <https://files.docking.org/catalogs/> alanına ait sayfa içeriğinin alt kısmından bir kesit.



Resim 52. İlaç adayı moleküllerin indirilebilmesi için gerekli “*download_list*” adlı dosya içeriğinin oluşturulmuş hali. (a) ZINC biyolojik veri tabanının statik dosyalarını barındıran <https://files.docking.org/> alanında bulunan klasör ve dosyaların yerel diske indirilmiş hali. (b) “*download_list*” adlı dosya içeriğinden bir kesit. (c) Kod betiğine ait “*log*” dosyalarının saklandığı klasör içeriğinden bir kesit.

Burada ifade edilen işlem adımlarının her biri için kullanılan dizin yapısını daha iyi anlamak için Şekil 17’yi incelemeniz önerilir.



Şekil 17. ZINC biyolojik veri tabanı statik dosyalarının yerel diske indirilmesi süreci için tasarlanan dizin yapısı ve elde edilen dosyalar. (a) Kök dizini ifade eder. (b) Kod betiğinin çalışma sürecinde belirtilen kök dizin içerisinde daha önceki adımlarda oluşturulmuş klasörleri ifade eder. (c) Kod betiğinin çalışma sürecinde belirtilen kök dizin için ilk defa oluşturulan klasörleri ifade eder. (d) Kod betiği işlemini tamamladığında elde edilen çıktı ve log kayıt dosyalarını ifade eder.

3.4.2. İlaç Adayı Molekül Havuzu Dosyalarının Elde Edilmesi Amacıyla Gerekli Ön Veri İşlemesi

Bu adımdaki işlemlere, bir önceki adımda elde edilen “*download_list*” dosya içeriğinde yer alan kayıtlardan, araştırma konumuzla ilgili veri barındıran ve “<*tranche_name*>.prot.txt” isim düzeninde arşivlenen her bir dosyanın okunarak içeriğinde yer alan verinin ait olduğu ZINC kataloğu altında toplanması ile başlanmıştır.

ZINC veri tabanında yer alan ve genel olarak bileşik olarak ifade edebileceğimiz dosyalar, “*tranches*” olarak adlandırılan dilimlere ayrılarak veri tabanı içerisinde saklanmaktadır. Dilimleme işlemi ise fiziksel özellik uzayında yer alan her bir veriye karşılık gelecek bir harflendirme yapılması yoluyla yapılmıştır (104).

Phyton programlama dilinde yer alan metin işlevleri kullanılarak her bir ZINC kataloğu için elde edilen içerikler öncelikle basit metin dosyası formatı olan “*txt*” türünde yerel diske yazılmıştır. Ardından bu dosyalar kod betiği iş akışındaki işlemlerin daha rahat yapılabilmesi için iki boyutlu tablo düzenine göre düzenlenen “*csv*” türüne çevrilmiştir.

Her iki dosya türü içindeki her bir kayıta yer alan veri alanlarına ait içerikler şu sıradadır: bir ilaç adayı molekülün SMILES metni, ZINC biyolojik veri tabanı özel erişim kimlik numarası, dilimleme işlemi için özel oluşturulan ve altı harfli olarak tasarlanan özellik sınıflama karakter dizisi, satın alınabilirlik değişkeni değeri ve reaktivite değişkeni değeri.

Son olarak bu süreçte elde edilen tüm dosyalar yerel diskte belirlenen kök dizin altındaki çalışma klasörünün altında “*raw_protomers*” adıyla kaydedilmiştir. Bu adıma ait sürecin daha iyi anlaşılması için Resim 53 incelenebilir.

Index of /catalogs/1

Parent Directory

chebi/

Index of /catalogs/1/chebi

Parent Directory

chebi.codemap.txt.gz

chebi.info.txt.gz

tranches/

Index of /catalogs/1/chebi/tranches

Parent Directory

AA/

AB/

AC/

AD/

AE/

AF/

AG/

AH/

AI/

AK/

BA/

BB/

BC/

BD/

BE/

Index of /catalogs/1/chebi/tranches/AC/

Parent Directory

chebi.codes

xaana-AC-mid

xaana-AC-ref

xaana-AC-mid.db2.gz

xaana-AC-mid.mol2.gz

xaana-AC-mid.pdbqt.gz

xaana-AC-mid.sdf.gz

xaana-AC-ref.db2.gz

xaana-AC-ref.mol2.gz

xaana-AC-ref.pdbqt.gz

xaana-AC-ref.sdf.gz

xaana-mid.db

xaana-mid.txt

xaana-ref.txt

xaana-prot.txt

xaab-AC-mid

xaab-AC-ref

xaab-AC-mid.db2.gz

xaab-AC-mid.mol2.gz

chebi.csv - LibreOffice Calc

999 GB Volume

Thesis_data

zinc_part

data

raw_promoters

a2z.csv

a2z.txt

acela.csv

acela.txt

acdisc.csv

acdisc.txt

acdiscnd.csv

acdiscnd.txt

achembloc.k.txt

aksci.csv

aksci.txt

alfa.csv

alfa.txt

alindabb.csv

alindabb.txt

aobious.csv

apollo.csv

apollo.txt

apollobb.csv

apollobb.txt

apollobio.csv

apollobio.txt

arkbb.csv

arkbb.txt

aronisbb.txt

asin.csv

asin.txt

bachem.csv

bachem.txt

bcp.csv

bcp.txt

bidebb.csv

bosche.csv

bosche.txt

capot.csv

capot.txt

cayman.csv

cayman.txt

cdibbe.csv

cdibbe.txt

chbr.txt

chbrbb.csv

chbrbb.txt

chbrbbe.csv

chbrbbe.txt

chbrbbp.csv

chbrbbp.txt

chbre.csv

chebi.csv

chebi.txt

chemaphor.csv

chemaphor.txt

chemimex.csv

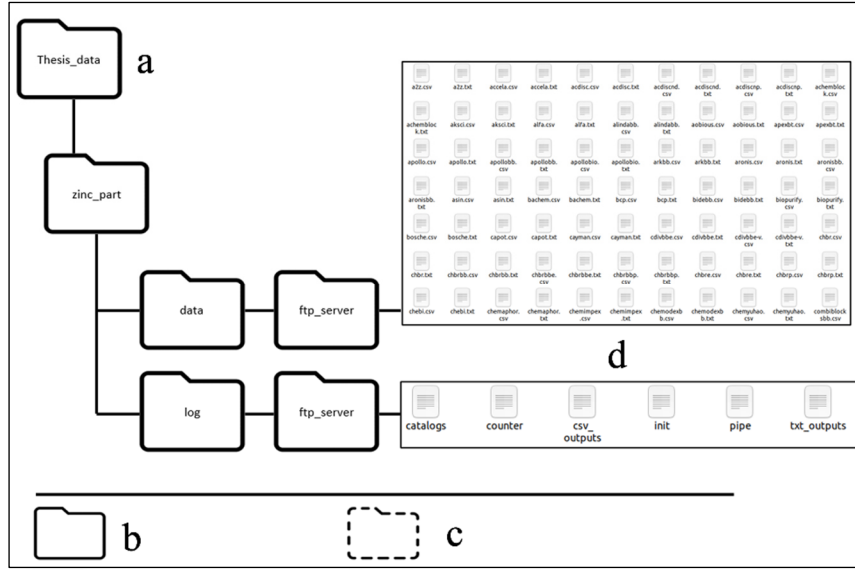
chemimex.txt

chemodexb.csv

chemodexb.txt

Resim 53. Yerel diske indirilen ham ilaç adayı molekülü dosyalarının oluşturulabilmesi için gerekli olan “*raw_promoters*” dizininin oluşturulma süreci.

Burada ifade edilen işlem adımlarının her biri için kullanılan dizin yapısını daha iyi anlamak için Şekil 18’i incelemeniz önerilir.



Şekil 18. İlaç adayı molekül havuzu dosyalarının elde edilmesi amacıyla gerekli ön veri işleme süreci için tasarlanan dizin yapısı ve elde edilen dosyalar. **(a)** Kök dizini ifade eder. **(b)** Kod betiğinin çalışma sürecinde belirtilen kök dizin içerisinde daha önceki adımlarda oluşturulmuş klasörleri ifade eder. **(c)** Kod betiğinin çalışma sürecinde belirtilen kök dizin için ilk defa oluşturulan klasörleri ifade eder. **(d)** Kod betiği işlemini tamamladığında elde edilen çıktı ve *log* kayıt dosyalarını ifade eder.

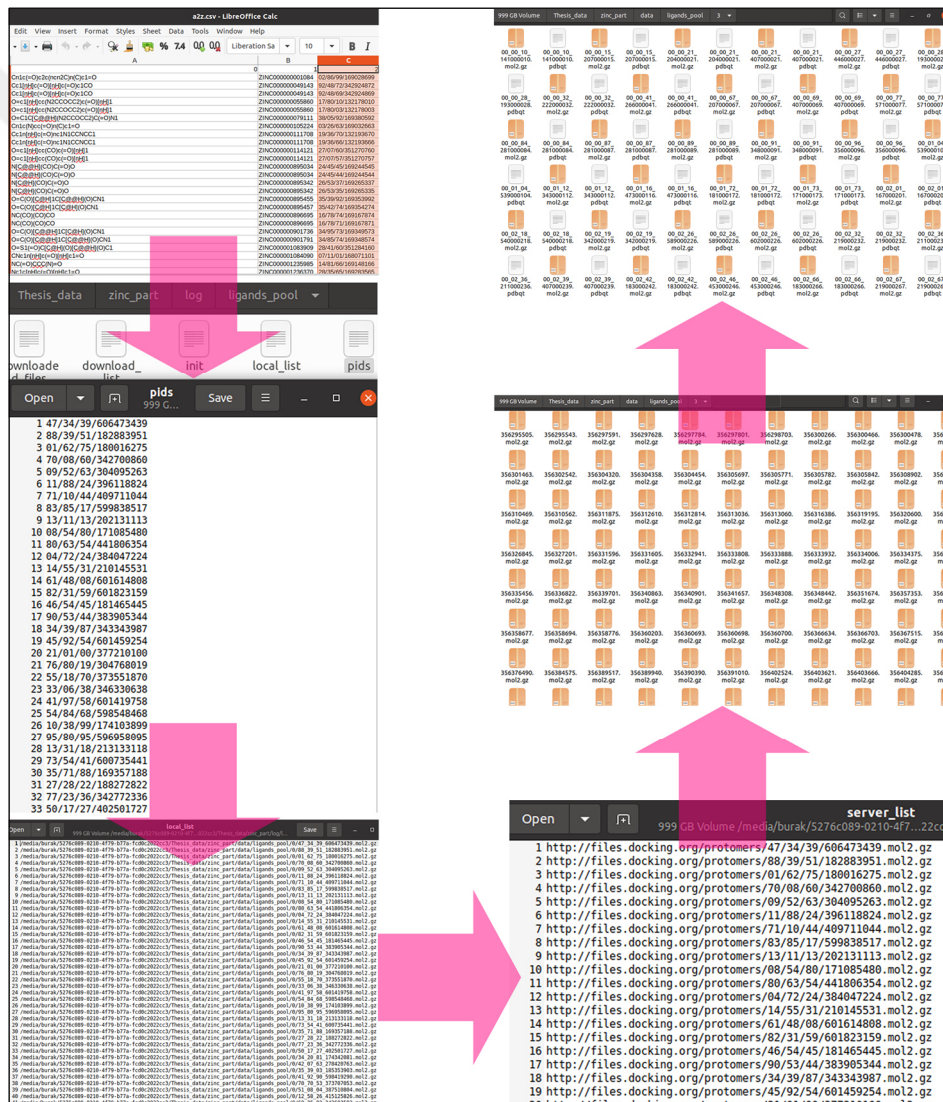
3.4.3. Sanal Tarama Analizleri İçin Gerekli İlaç Adayı Molekül Dosya Havuzunun Oluşturulması

Bu kısımda bir önceki adımda her bir ZINC kataloğu için elde edilmiş olan ve “*raw_protomers*” klasörü altında yer alan “*csv*” uzantılı dosyalar kod betiğinin girdi verileri olarak kullanılmıştır. Her bir girdi dosyasının üç numaralı veri alanına karşılık gelen kısımlarda yer alan kayıtlar her bir katalog için okunmuştur. Bu kayıtlardan tekil olanlar saklanarak elde edilen veriler “*protomer_ids*” kavramına karşılık gelecek ve anlamı “*protomer_id*” olan “*pids*” adlı log kayıt dosyasına aktarılmıştır.

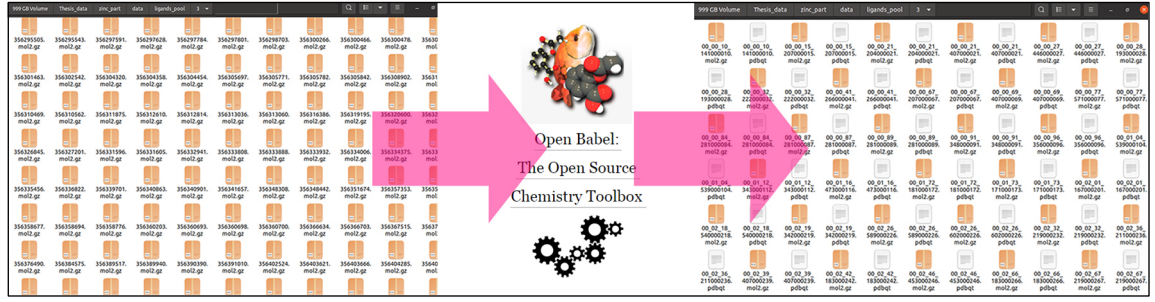
Ardından “*pids*” dosyasındaki her bir kayıt için http://files.docking.org/protomers/<protomer_id>.mol.gz formatında düzenlenen istek sunucuya gönderilmiş ve sunucudan cevap olarak gelen dosyalar “*ligands_pool*” klasörü altında “<protomer_id>.mol.gz” dosya adı düzeninde kayıt edilmiştir. Dosyalar yerel diske kaydedilirken toplam dosya adedinin milyonları bulacak seviyede olması nedeniyle bilgisayar kapasitesini çok zorlamamak adına alt klasörleme işlemi yapılmıştır. Böylece toplam 1 363 175 adet olan dosyalar, her bir klasörde maksimum 100 000 dosya tutulacak şekilde alt klasörlere ayrılarak yerel diske kaydedilmiştir.

Ardından “*mol.gz*” uzantılı sıkıştırılmış dosya formatından oluşan her bir dosya “*Phyton*” programlama dili ve “*Open Babel*” yazılımı yardımıyla sanal tarama analizleri için kullanılacak yardımcı yazılımın girdi dosya formatı olan “*pdbqt*” türüne dönüştürülmüştür (105, 106). Bu dönüşümler yapılırken “*protomer_id*” verisi farklı olan çoğu dosyanın aynı ZINC biyolojik veri tabanı erişim kimliği verisine sahip olduğu görülmüştür. Kısacası yerel diskte oluşturulan ilaç adayı molekül havuzu kaynak dosyaları içerisinde aynı kimyasal bileşiğe karşılık gelen dosyalar olduğu gözlenmiştir.

Bu adıma ait sürecin daha iyi anlaşılması için Resim 54 ve 55 incelenebilir.

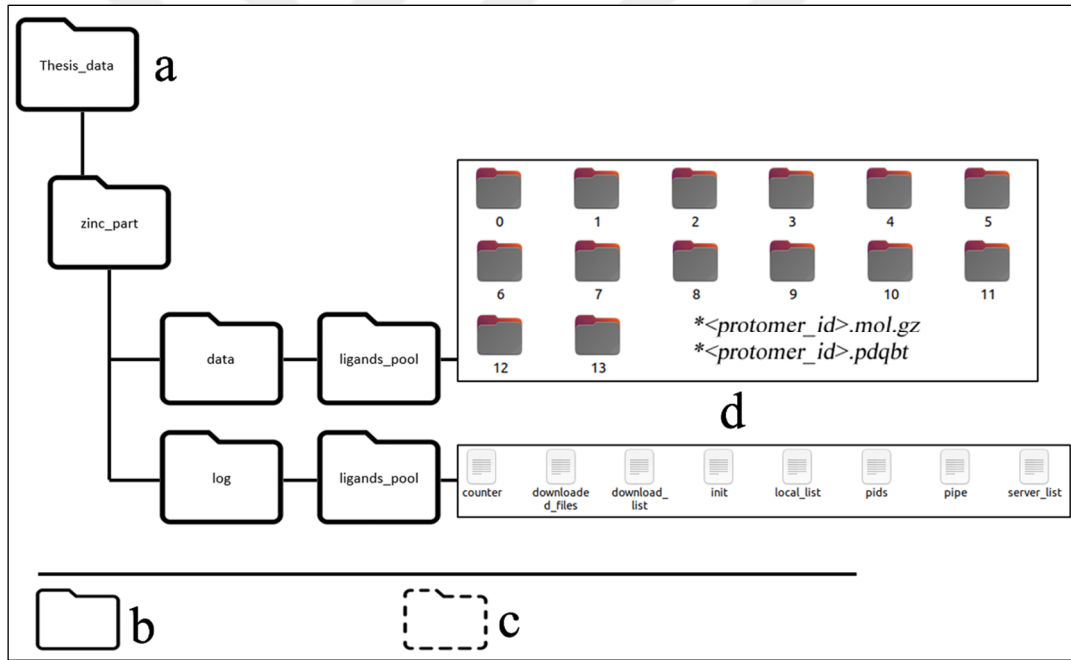


Resim 54. Sıkıştırılmış arşiv dosyası “*mol.gz*” türündeki üç boyutlu “*ready to dock*” düzenindeki dosyaların yerel diske indirilmesi süreci



Resim 55. Yerel diske indirilmiş olan “*mol.gz*” formatındaki üç boyutlu “*ready to dock*” düzenindeki dosyaların ligand “*pdbqt*” formatına dönüşüm sürecini temsil eden bir kesit.

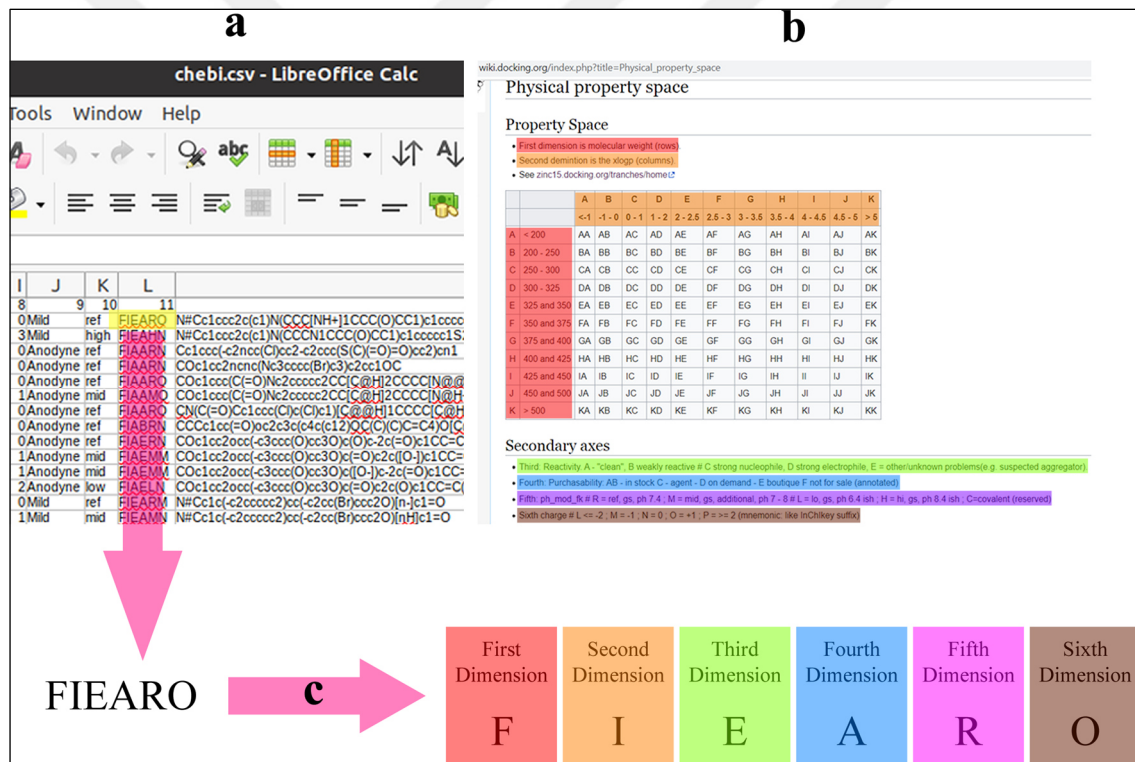
Burada ifade edilen işlem adımlarının her biri için kullanılan dizin yapısını daha iyi anlamak için Şekil 19’u incelemeniz önerilir.



Şekil 19. Sanal tarama analizleri için gerekli ilaç adayı molekül dosya havuzunun oluşturulması süreci için tasarlanan dizin yapısı ve elde edilen dosyalar. (a) Kök dizini ifade eder. (b) Kod betiğinin çalışma sürecinde belirtilen kök dizin içerisinde daha önceki adımlarda oluşturulmuş klasörleri ifade eder. (c) Kod betiğinin çalışma sürecinde belirtilen kök dizin için ilk defa oluşturulan klasörleri ifade eder. (d) Kod betiği işlemini tamamladığında elde edilen çıktı, rapor, tanımlayıcı veri, grafik ve *log* kayıt dosyalarını ifade eder.

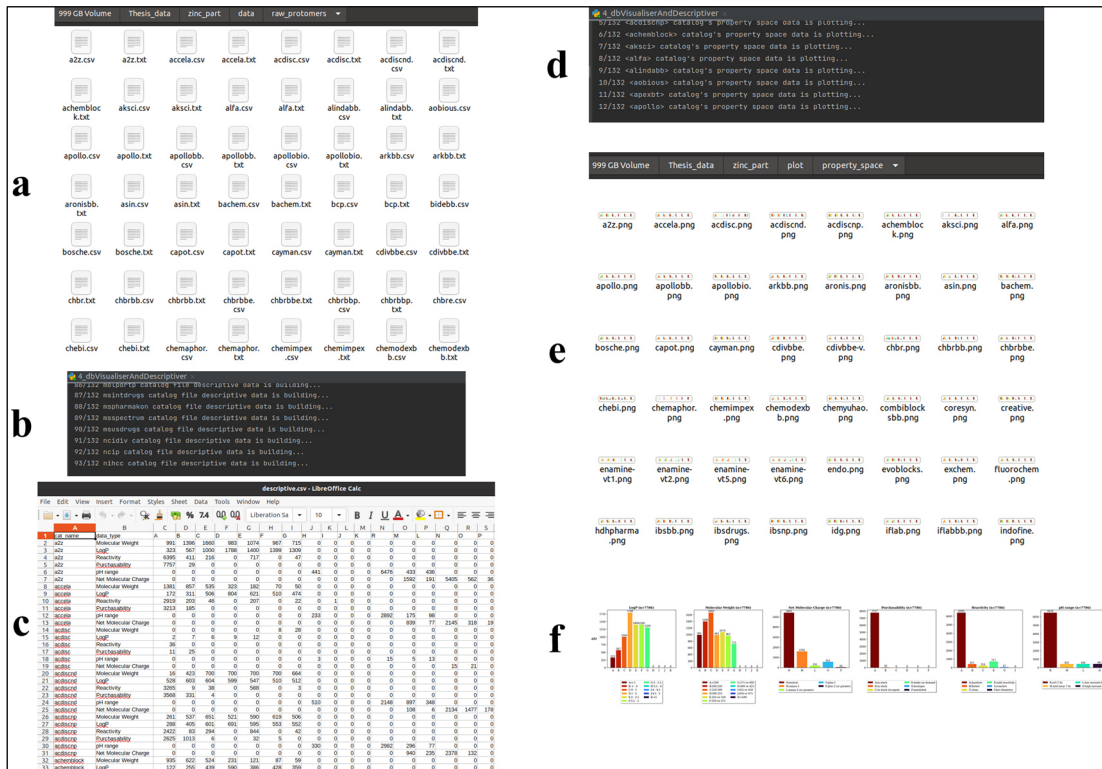
3.4.4. İlaç Adayı Molekül Dosya Havuzunu Oluşturan ZINC Katalog Verilerine Ait Fiziksel Özellik Uzaı Veri Sıklıklarının Hesaplanması ve Grafiklerinin Çizdirilmesi

Bu kısımda daha önceki işlem adımlarında oluşturulan “*raw_protomers*” klasörü içerisinde yer alan dosyalar girdi olarak kullanılmıştır. Kullanılan girdi dosyaları “*<zinc_katalog_adı>.csv*” şeklinde adlandırılmış olup her bir dosya ZINC kataloğuna ait fiziksel özellik verisi içermektedir. Her bir katalog verisi, içerisinde fiziksel özellik uzayı verilerini daha rahat elde edebilmek amacıyla düzenlenmiş “*özel bir veri alanına*” sahiptir. Bu özel veri alanı içerisinde, altı harften oluşan özel bir kodlama verisi bulunmaktadır. Resim 56’da bu kodlamanın nasıl elde edildiği ve her bir karakterin karşılık geldiği veri türü ile alabileceği değerler açık bir şekilde betimlenmiştir.

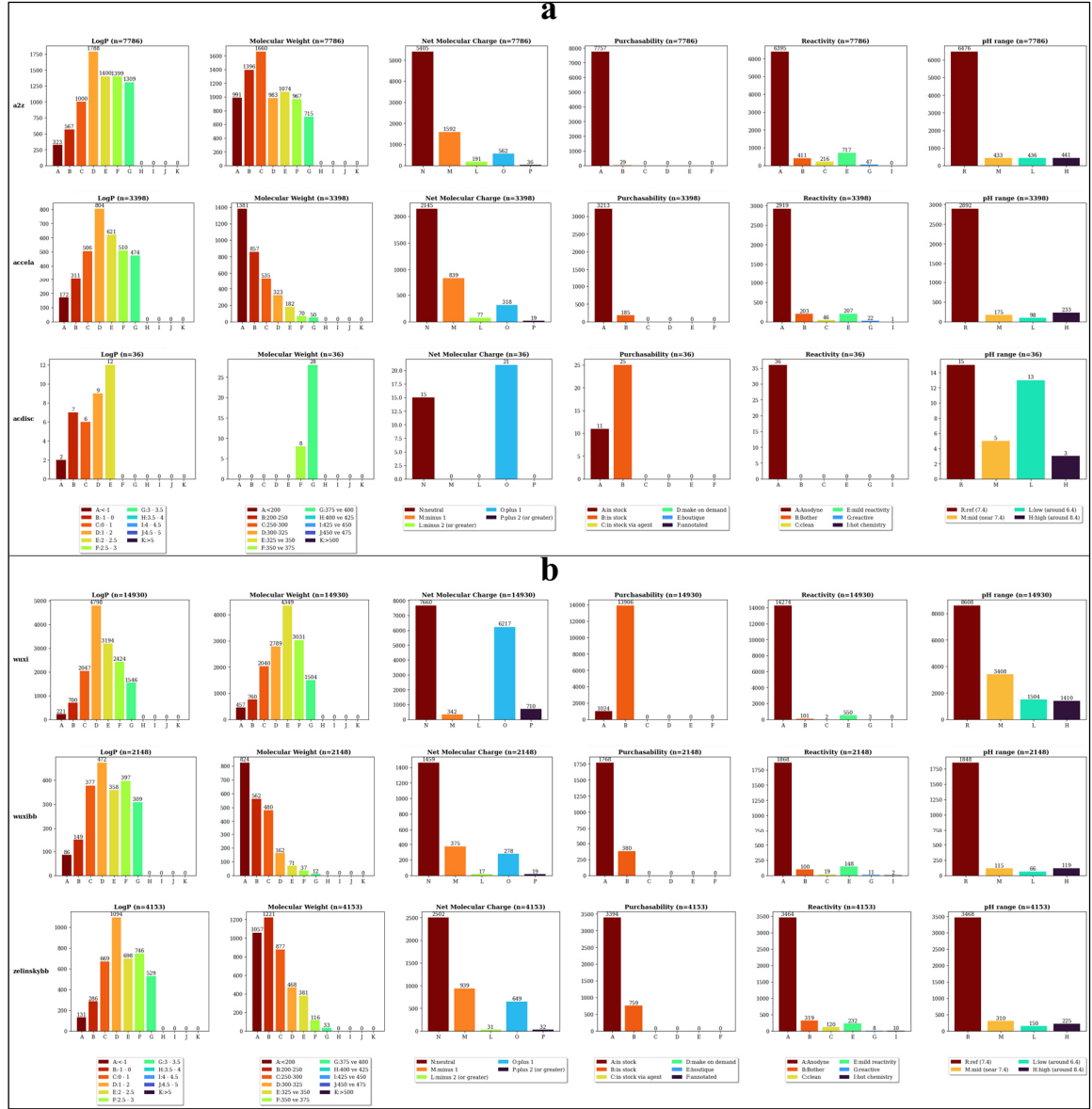


Resim 56. Kod betiği ile elde edilen ZINC kataloglarına ait fiziksek özellik uzayı sıklık dağılım verileri üzerinden, tüm kataloglara ait veriyi içeren “*descriptives.csv*” dosyasının oluşturulması sürecinde yapılan işlem mantığını betimleyen den bir kesit. **(a)** Örnek bir katalog veri girdi dosyası içeriği ve kullanılan özel veri alanını vurgulayan bir kesit. **(b)** Grafik çıktılarında kullanılan lejant ile özel veri alanında kullanılan harf kodlamasında her bir harfin karşılık geldiği değeri gösteren bir kesit. **(c)** Özel veri alanında yer alan altı harfli kodlamanın açılımını ve temsil ettiği değişken boyutlarını betimleyen bir kesit.

Bu kısım için yazar tarafından hazırlanan kod betiği, açıklaması verilen bu özel veri alanındaki değerleri işlemiştir. Böylece her bir ZINC kataloğu için altı temel özelliğe ait değişkenlerin ilk olarak sıklık dağılım verisi elde edilmiştir. İkinci olarak ise elde edilen bu sıklık dağılım verileri girdi olarak kullanılmış ve yine bu verileri temsil eden sıklık dağılım grafikleri de başarıyla çizdirilmiştir. Bu adımlara ait sürecin daha iyi anlaşılması için Resim 57 ve 58 incelenebilir.

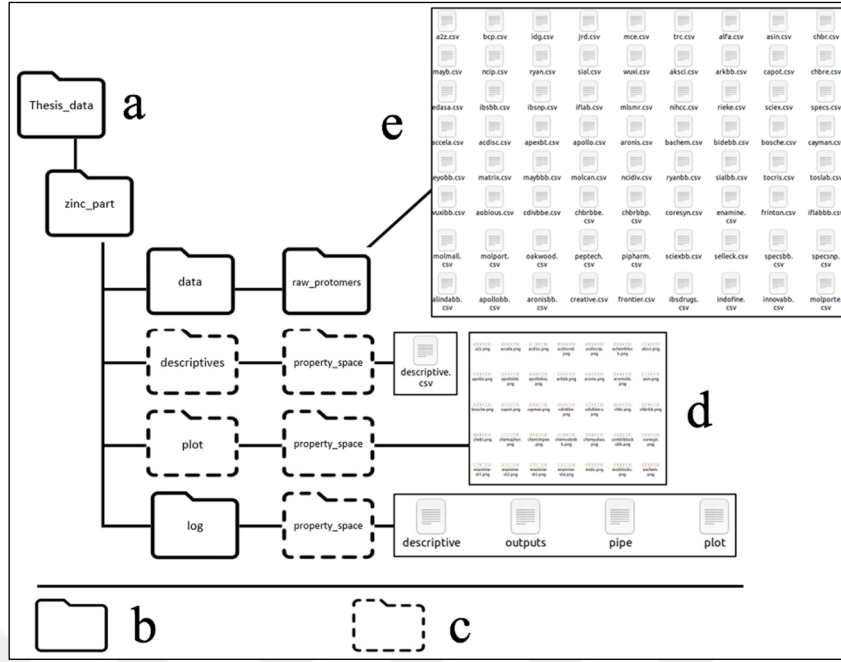


Resim 57. Kod betiği ile elde edilen ZINC kataloglarına ait fiziksek özellik uzayı sıklık dağılım verileri üzerinden, sıklık dağılım grafiklerinin elde edilmesi sürecindeki işlem adımlarını temsil eden bir kesit. (a) Kod betiği için gerekli girdi dosyalarından bir kesit. (b) Girdi dosyalarından sıklık verilerini oluşturan kod betiğinin çalışma sürecinden bir kesit. (c) Elde edilen sıklık veri dosyası içeriğinden bir kesit. (d) Elde edilen sıklık veri dosyası üzerinden grafik çıktıları oluşturan kod betiğinin çalışma sürecinden bir kesit. (e) Elde edilen grafik çıktı dosyalarının yerel diskteki konumundan bir kesit. (f) Elde edilen grafik çıktı dosyalarından örnek bir gösterim.



Resim 58. ZINC veri tabanını oluşturan her bir kataloğa ait fiziksel özellik uzayı verilerini temsil eden sıklık dağılım grafiği çıktılarından bir kesit. **(a)** Katalog isimlerine göre alfabetik olarak sıralanmış ilk üç kataloğa ait dağılım grafik çıktıları. **(b)** Katalog isimlerine göre alfabetik olarak sıralanmış son üç kataloğa ait dağılım grafik çıktıları.

Burada ifade edilen işlem adımlarının her biri için kullanılan izin yapısını daha iyi anlamak için Şekil 20'yi incelemeniz önerilir.



Şekil 20. “İlaç Adayı Molekül Dosya Havuzunu Oluşturan ZINC Katalog Verilerine Ait Fiziksel Özellik Uzayı Veri Sıklıklarının Hesaplanması ve Grafiklerinin Çizdirilmesi” süreci için tasarlanan dizin yapısı ve elde edilen dosyalar. **(a)** Kök dizini ifade eder. **(b)** Kod betiğinin çalışma sürecinde belirtilen kök dizin içerisinde daha önceki adımlarda oluşturulmuş klasörleri ifade eder. **(c)** Kod betiğinin çalışma sürecinde belirtilen kök dizin için ilk defa oluşturulan klasörleri ifade eder. **(d)** Kod betiği işlemini tamamladığında elde edilen çıktı, tanımlayıcı veri, grafik ve log kayıtlarını ifade eder. **(e)** Kod betiğinin çalışması için gerekli girdi dosyasını ifade eder.

3.5. Sanal Tarama Analizi Kısımı

Bu kısımda, bu noktadan sonra kısaca “vina” olarak ifade edilecek olan “Autodock Vina”(76, 77) yazılımı desteği ile yürütülen sanal tarama süreci için yapılan çalışmalara değinilecektir. Sanal tarama (82) süreci için yapılan ilk işlem “vina”(76, 77) yazılımının kullanılacak olan sunucuya yüklenmesi işlemiydi. Bu amaç için “Anaconda”(62) dağıtıcısından ulaşılabilen “vina”(77) paketinin kurulumu ilgili sunucuya yapılmıştır. Ardından işleyen süreç ise sırasıyla; “vina”(76, 77) yazılımı için gerekli girdi dosyalarının hazırlanması, oluşacak çıktı dosyalarının da düzgün şekilde arşivlenmesi ve raporlanması için de gerekli kod altyapısının hazırlanmasıdır.

Bu bölüme ait gerekli dosyaların elde edilebilmesi için “Phyton” programlama dili ve bu dile ait çeşitli paketlerin desteğiyle birlikte yazar tarafından oluşturulan kod betikleri kullanılmıştır (60). Kod betiklerinin yaptıkları işlemlere ait kısa açıklamalar tabloda verilmiştir. Tablo 9’da yer alan her bir kod betiğinin takip ettiği iş akış sürecinde yapılan

işlemler ise bu bölüme ait takip eden alt bölüm başlıkları halinde detaylı olarak ifade edilmeye çalışılmıştır.

Tablo 9. Sanal tarama kısmı için oluşturulan kod betikleri ve kısa açıklamaları

Kod betiğinin adı	Yaptığı işleme dair kısa açıklama
<i>0_vsReceptorSelectorAndPreparer.py</i>	Reseptör adayı dosyalar arasından belirtilen kriterlere uyan bir tanesinin seçimi ve analiz için uygun hale getirilmesi
<i>1_vsAccuracyCalculator.py</i>	Sanal tarama analizi için daha önceden elde edilmiş ilaç adayı moleküllerden küçük bir örneklem seçilmesi ve elde edilen sonuçlar ile başka bir sanal tarama analizi yapan kaynağın, aynı dosyalar üzerinden elde edeceği sonuçları ile karşılaştırılması
<i>2_vsJobsRunner.py</i>	Elimizdeki tüm ilaç adayı molekülleri kullanan bir sanal tarama analizi sonuçlarının elde edilmesi
<i>3_vsResultsReporter.py</i>	Elimizdeki tüm ilaç adayı molekülleri kullanan bir sanal tarama analizi sonuçları arasından en iyi bulgulara sahip moleküllerin raporlanması

3.5.1. Sanal Tarama Analizi İçin Aday Reseptörler Arasından Dosya Seçimi

Bu kısımda önceki adımlarda tespit edilen reseptör adayları içerisinde (receptor_candidates.csv) araştırma kriterlerimize uygun bir tanesinin rastgele olarak seçilmesi işlemi gerçekleştirilmiştir. Tek reseptör aday seçmemiz çalışmanın bir kısıtı olarak yorumlanabilir. Ancak, bu tez çalışmasında oluşturmaya çalıştığımız iş akış süreçleri kullanılarak birden çok reseptör dosyasının kullanıldığı çalışmalar planlanabilir.

Çok fazla ilaç adayı molekül dosyası (milyonlarca) kullanılarak planlanan sanal tarama analizlerinin, hızlı sonuçlandırılabilmesi için güçlü işlem gücüne sahip sunucu bilgisayarları gerekmektedir. Kişisel bilgisayarların işlem gücü kapasitesi bu alanda oldukça yetersiz kalmaktadır. Bu amaçla yazar, bu çalışmayı belirlenen zamanda tamamlayabilmek için, tez çalışmasını bünyesinde yürüttüğü KTÜ Biyoistatistik ve Tıp Bilişimi ABD “Kanserde İmmünoterapi Etkinliğini Artırmak İçin Makine Öğrenmesine Dayalı Gen Keşfi ve İlaç Yeniden Konumlandırma Platformunun Geliştirilmesi” adlı TÜBİTAK projesinden sağlanan TRUBA sunucu bilgisayarlarına erişim desteği almıştır.

Benzer şekilde, büyük boyutlu sanal tarama analizlerinin tamamlanması zaman açısından da uzun süreler alabilmektedir. Bu tez çalışması yürütülürken de hem takip edilmesi gereken bir zaman çizelgesi hem de uyulması gereken bir zaman sınırı

bulunmaktaydı. İlgili konu başlığında detayı verilecek olan ve tek reseptör dosyası üzerinden yapılan sanal tarama analizlerimiz TRUBA kaynak sunucu bilgisayarları ile çalışılmasına rağmen yaklaşık altı aylık bir sürede ancak tamamlanabilmiştir.

Neden tek reseptör dosyasının seçildiği bir çalışma planladığımızın daha net anlaşılması için aşağıda verilen literatür destekli açıklamaların da önem arz ettiğini düşünmekteyiz. Literatürde, sanal tarama analizlerinin kullanılacak dosya sayısına göre toplamda kaç saatlik bilgisayar işlem gücüne ihtiyaç duyacağının hesaplamasına dair kesin sonuç veren bir çalışma tespit edilememiştir. Bu durumun sanal tarama analizleri içerisinde kontrol edilmesi gereken birçok farklı etmenin olduğu dikkate alındığında anlaşılabilir olduğunu düşünmekteyiz. Ancak biz çalışmamız için en azından genel çerçeveyi çizebilmek adına literatürde yer alan örnek bir çalışmadaki verileri dikkate aldık (107). Bu çalışmada yaklaşık 15 000 000 “*docking*” işlemi (*docking* işlemi=reseptör dosya sayısı x ilaç adayı molekül sayısı) için yaklaşık 500 000 işlemci saati kullanıldığı ifade edilmiştir (107). Örnek olarak bu çalışmanın seçilme nedeni ise “*docking*” işlemleri için, çalışmamızdaki gibi “*vina*” yazılımının kullanılmış olmasıdır (76, 77).

Çalışmamız için kullandığımız TRUBA kaynağı ise bize 420 000 işlemci saatlik bir kaynak tahsis etmiştir. Örnek çalışmada 260 000 ilaç adayı dosyaya karşılık 58 adet reseptör dosyası kullanılmıştır (107). Bu durumda benzer koşullar altında, çalışmamızda ise yaklaşık 1 400 000 ilaç adayı moleküle karşılık 10 adet reseptörün kullanılabileceği düşünülmüştür. Ancak kaynak makalede yazılımın en önemli parametresi olan ve işlemci gücünü direkt etkileyen “*exhaustiveness*” parametresinin varsayılan değer olan sekiz olarak alındığı görüldü. Biz ise çalışma sonuçlarının doğruluğunun ve kesinliğinin daha iyi çıkması adına “*exhaustiveness*” parametresini daha üst bir değer olan 32 olarak almayı tercih ettik. Yerel bilgisayarda yaptığımız denemelerde “*exhaustiveness*” parametre değerinin sekizden 32 ‘ye çıkmasının analizin sonuçlanma süresinin yaklaşık olarak beş kat daha uzamasına neden olduğu gözlemlendi. Bu ön bulguyla birlikte tahmini bir hesap yaparak iki farklı reseptör kullanılabileceği düşünüldü ve gerekli kod betiği buna göre düzenlendi. Ancak tek reseptör için başlayan süreçte bize tahsis edilen kaynağın neredeyse tamamının (%97.6) kullanıldığı görüldü. Bu nedenle çalışma tek reseptör adayı dosya üzerinden sonuçlandırılmıştır.

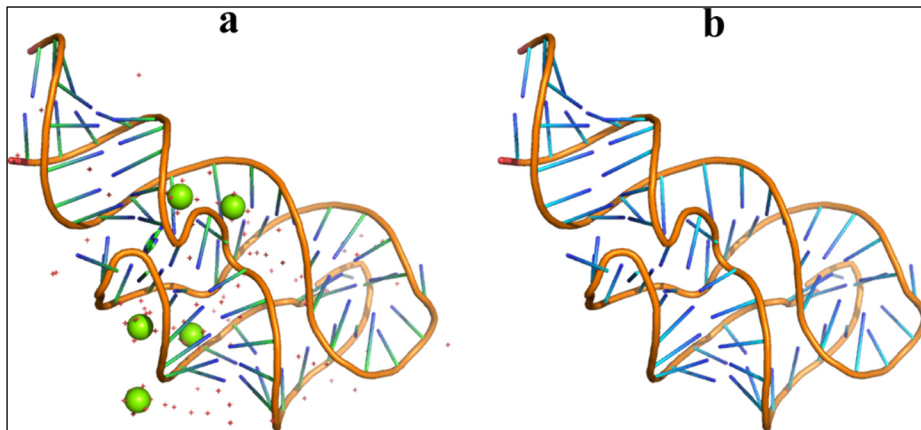
Yukarıda açıkça ifade edilen bilgisayar işlemci gücü ve zaman kısıtlarından dolayı bir önceki adımda elde edilen reseptör adayı dosyaları içerisinde tek bir “*pdb*” dosyası olan organizmalar seçilmiştir. Seçilen bu organizmalara karşılık gelen “*pdb*” dosyaları

içerisinden bir tanesi rastgele olarak seçilmiştir. Bu dosya ise protein veri bankası erişim kodu “1Y26”(108) ve kaynak genomu “*V. vulnificus*” olan dosyaya aittir. Bu dosya mevcut haliyle “vina” yazılımına girdi olarak verilmeye uygun değildir (76, 77). Dosya üzerinden öncelikle “temizleme” yapılması gerekmektedir.

Temizlemeden kastedilen;

“Burada verilen PDB dosyasına ait satırlar içerisinde protein, ligand su bileşiğindeki oksijen atomlarını içeren kısımlar olup olmadığını kontrol ediyoruz. İlk olarak, sadece ATOM anahtar kelimesiyle başlayan satırlara karşılık gelen protein atomlarını seçiyoruz. Her bir protein zinciri, TER ile başlayan bir satır ile sonlandırılır. Seçimlerimizi TER ile başlayan satıra denk gelince sonlandırıyoruz (109).”

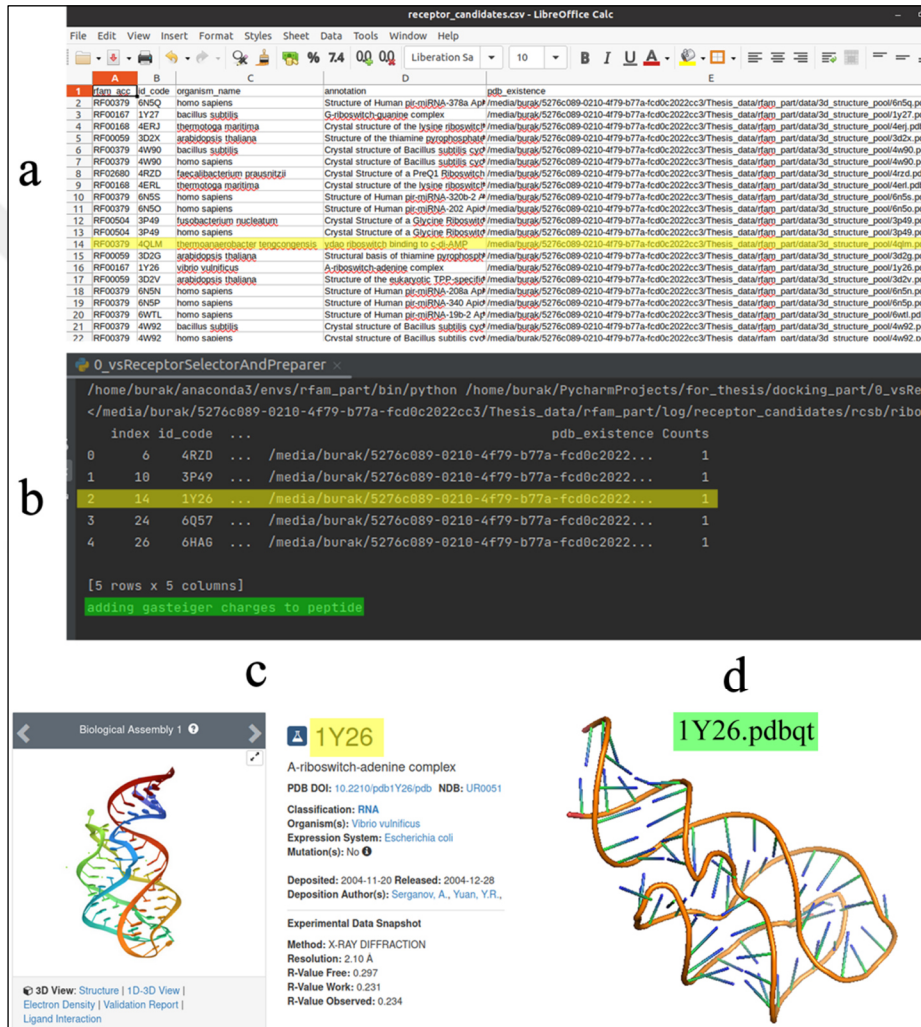
Temizleme işlemi yapıldıktan sonra temizlenmiş protein dosyasının (1y26_protein.pdb) “vina”(76, 77) yazılımının girdi formatına uygun hale dönüştürülmesi gerekmektedir. Bu dönüşüm işlemi için ise “MglTools”(75, 110) yazılımının kurulum dizini altındaki “VSUtilities24” klasörü içerisinde yer alan “prepare_receptor4.py” kod betiği kullanılarak yapılmıştır (75). Bu adımlara ait sürecin daha iyi anlaşılması için Tablo 10 ile Resim 59 ve Resim 60 incelenebilir.



Resim 59. Sanal tarama analizi için seçilen aday reseptör dosyasında temizleme işlemi yapılma süreci. (a) Reseptör dosyasının temizlenmeden önceki hali (1Y26.pdb) (b) Reseptör dosyasının temizlenmiş hali (1Y26_protein.pdb)

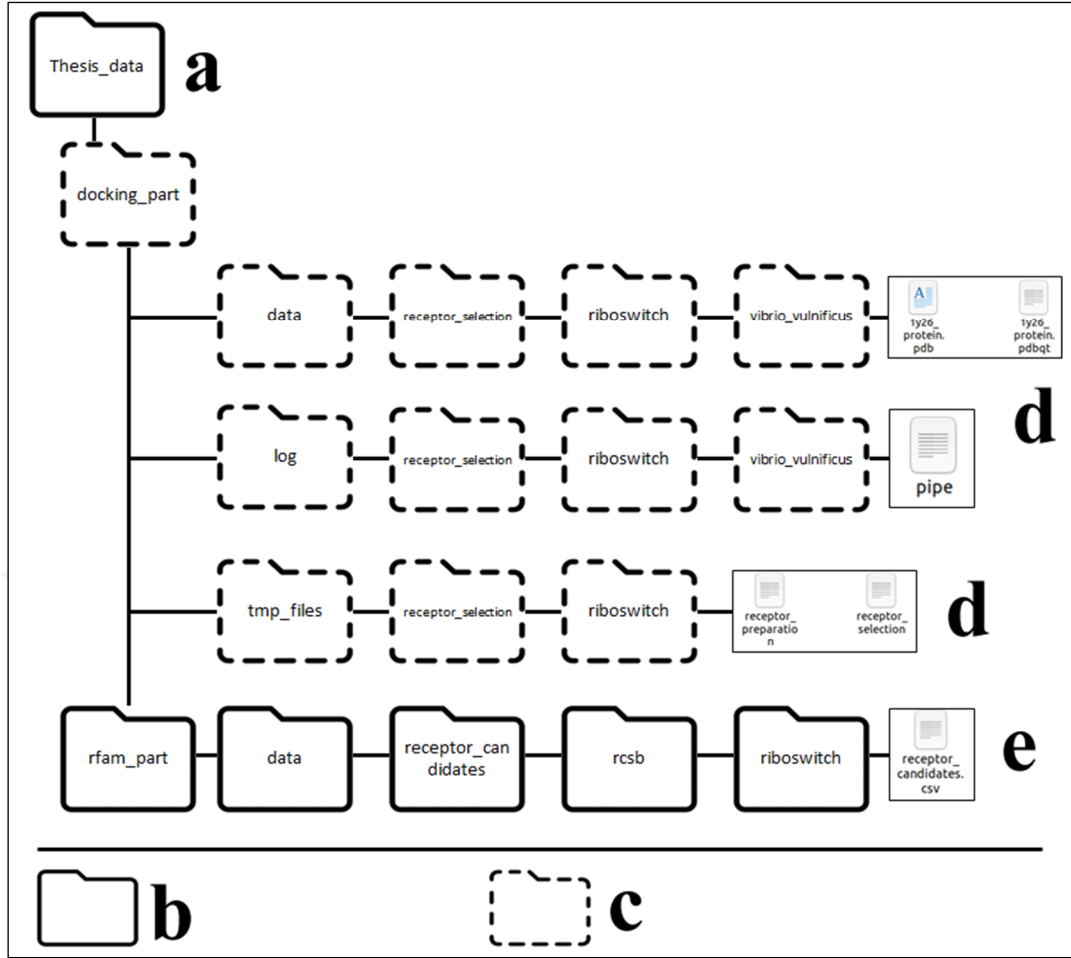
Tablo 10. Reseptör adayları dosyaları içerisinde tekil dosyaya sahip olan kayıtlar

Pdb Erişim Kimliği	Kaynak Genom
4RZD (111)	<i>Faecalibacterium prausnitzii</i>
3P49 (112)	<i>Fusobacterium nucleatum</i>
1Y26 (113)	<i>V. vulnificus</i>
6Q57 (114)	<i>Streptococcus mutans</i>
6HAG (115)	<i>Proteobacteria (E. coli)</i>



Resim 60. Sanal tarama analizi için aday reseptör dosyalarından seçim yapılması ve seçilen dosya içeriğinin analiz için hazır hale getirilmesi süreci. **(a)** Reseptör adayları dosya içeriğinden bir kesit. **(b)** Kod betiği çalışma sürecinden bir kesit. **(c)** Seçilen reseptör dosyasının işlenmemiş ham halinden bir kesit. **(d)** Seçilen reseptör dosyasının işlenmiş ve analiz için dönüştürülmüş saf halinden bir kesit.

Burada ifade edilen işlem adımlarının her biri için kullanılan dizin yapısını daha iyi anlamak için Şekil 21'i incelemeniz önerilir.

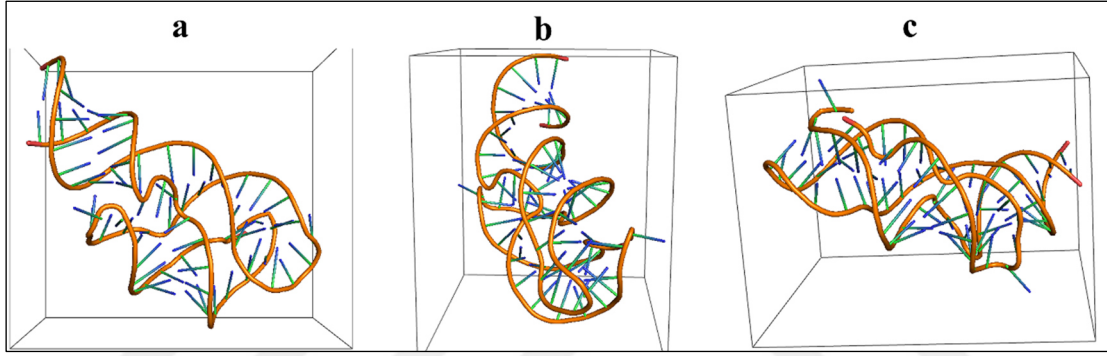


Şekil 21. Sanal tarama analizi için aday reseptör dosyalarından seçim yapılması ve seçilen dosya içeriğinin, analiz için hazır hale getirilmesi süreci için tasarlanan dizin yapısı ve elde edilen dosyalar. **(a)** Kök dizini ifade eder. **(b)** Kod betiğinin çalışma sürecinde belirtilen kök dizin içerisinde daha önceki adımlarda oluşturulmuş klasörleri ifade eder. **(c)** Kod betiğinin çalışma sürecinde belirtilen kök dizin için ilk defa oluşturulan klasörleri ifade eder. **(d)** Kod betiği işlemini tamamladığında elde edilen çıktı, istatistiksel veri ve *log* kayıt dosyalarını ifade eder. **(e)** Kod betiğinin çalışması için gerekli girdi dosyasını ifade eder.

3.5.2. Sanal Tarama Analizi İçin Gerekli Girdi Dosyalarının Oluşturulması

Bu kısımda “vina”(76, 77) yazılımının düzgün çalışabilmesi için gerekli girdi dosyalarının oluşturulma sürecine değinilecektir. Sanal tarama analizlerinin yapılacağı “vina”(76, 77) yazılımı girdi dosyası olarak üç türde dosyaya ihtiyaç duymaktadır. Bunlardan birincisi olan *reseptör* dosyası, bir önceki başlıkta belirlenen şekilde elde edilmiştir. İkinci girdi dosyası ise ilaç adayı molekül dosyalarıdır. Bu dosyalar da Başlık 3.4.3’te anlatılan şekilde elde edilmiştir. Bu kısım için eksik kalan “vina box” olarak adlandırılan girdi dosyasının oluşturulması için gerekli kod betiği düzenlenmiştir. Sanal

tarama analizleri için kullanmayı taahhüt ettiğimiz “*blind docking*”(74) yöntemi için gerekli alan hesaplaması bahsi geçen üçüncü girdi dosyası ile yapılmaktadır. “*Blind docking*”(74) yöntemi reseptör dosyasının üç boyutlu ekseninde kapladığı alanın tamamını kapsayacak biçimde dikdörtgenler prizması şeklinde kutu boyutlarının ve X, Y, Z merkez koordinatlarının tespiti işlemlerini ifade eder. Literatürde, bu işi yapığını iddia eden çeşitli kod betikleri yer almaktadır. Ancak biz burada “*Python*”(60) programlama dilinin “*Rdkit*”(116) paketi altında yer alan çeşitli kütüphanelere ait fonksiyonları kullanarak gerekli koordinat nokta değerlerini kendimiz hesaplamayı tercih ettik (Resim 61). Bahsi geçen paket “*Pypi*”(63) üzerinden indirilerek ilgili dizin altına yüklenmiştir.



Resim 61. Sanal tarama analizi için seçilen aday reseptörün kutu boyutu girdi değerlerinin hesaplanma süreci. (a) X eksenindeki görünüm (b) Y eksenindeki görünüm (c) Z eksenindeki görünüm

3.5.3. Sanal Tarama Analizi Öncesinde “*Docking*” Başarısının ve Doğruluğunun Tespiti İçin Ön Analiz Yapılması

Önceki adımlarda yapılan açıklamalardan da anlaşılacağı üzere sanal tarama analizi hem oldukça zaman alan hem de ciddi anlamda bilgisayar işlemci gücü gerektiren bir işlemdir. Bu nedenle sanal tarama analizlerini başlatmadan önce elde edeceğimiz sonuçların tutarlı ve literatürde kabul görmüş benzer çalışmalarla uyumlu olduğunun tespit edilmesi bir gerekliliktir. Bu amaçla küçük çaplı bir ön sanal tarama analizi yapılması kararlaştırıldı.

Ön analizde kullandığımız kod betiğinin iş akışını şöyle özetleyebiliriz. İlk olarak çalışmamızda kullandığımız “*vina*”(76, 77) yazılımını kullanan ve literatürde yer alan bir çevrimiçi “*docking*” sunucusu araştırıldı. Yapılan literatür taraması sonucunda “*webina*” sitesi ön analizde kullanılmak üzere seçildi (117).

Ardından bu süreç için oluşturulan dizin yapısında “*local_run*” adıyla oluşturulan klasör kendi analizlerimizi, “*web_run*” adıyla oluşturulan klasör ise seçilen sunucuya ait analizleri tanımlayacak şekilde oluşturuldu. Analiz sonuçlarının karşılaştırılması için ICC istatistiği tercih edildi (91, 118). Bu istatistiğin yeterli sonuç vermesi için örneklem sayısı 50 olarak seçildi (119).

Ardından ilaç adayı dosya havuzundan (*ligands_pool* klasörü) rastgele 50 tanesi seçilerek “*ligands_list*” dosyasına yazıldı. Sonrasında önceki adımlarda seçilen ve gerekli dönüşümü yapılan reseptör dosyasının konum bilgisi ve bu reseptör dosyası için hesaplanan “*vina*” kutusu (*vina box*) koordinat merkezi değerleri hafızaya alındı. Hafızaya alınan bu değerler ile seçimi yapılan 50 adet ilaç adayı dosyanın konum bilgileri harmanlanarak “*vs_inputs*” adlı dosya içeriği oluşturuldu. Oluşturulan bu dosya içeriği “*vina*”(76, 77) yazılımının düzgün çalışabilmesi için gerekli tüm girdi parametrelerini içerecek şekilde düzenlendi.

Ardından “*vs_inputs*” dosyası içerisinde yer alan girdiler kullanılarak yerel disk üzerinden ön sanal tarama analizi gerçekleştirildi. Analize ait sonuç dosyaları “*local_run*” klasörü altında kaydedildi. Aynı girdiler kullanılarak bu kez “*webina*”(117) sunucu bilgisayarı üzerinden ön sanal tarama analizi gerçekleştirildi. Bu analize ait sonuç dosyaları ise “*web_run*” klasörü altında kaydedildi. Son olarak ise her iki ön analiz yöntemi ile elde edilen sonuç dosyaları okunarak içlerinden en iyi üç “*docking*” konformasyonuna sahip moleküllerin üç boyutlu konum verileri “*csv*” formatında elde edilmiştir. Üç boyutlu konum verilerini barındıran ve adları sırasıyla “*points_icc_data.csv*” ile “*centroids_data.csv*” olan iki adet dosya oluşturulmuştur. Birinci dosya içeriğinde tüm üç boyutlu konum nokta koordinatları verisi yer almaktadır. İkinci dosya içeriğinde ise sadece her bir konformasyonu oluşturan ilaç adayı molekülün ağırlık merkezlerine ait üç boyutlu konum nokta koordinatları verisi yer almaktadır.

Bu adımlara ait sürecin daha iyi anlaşılması için Resim 62-64 incelenebilir.



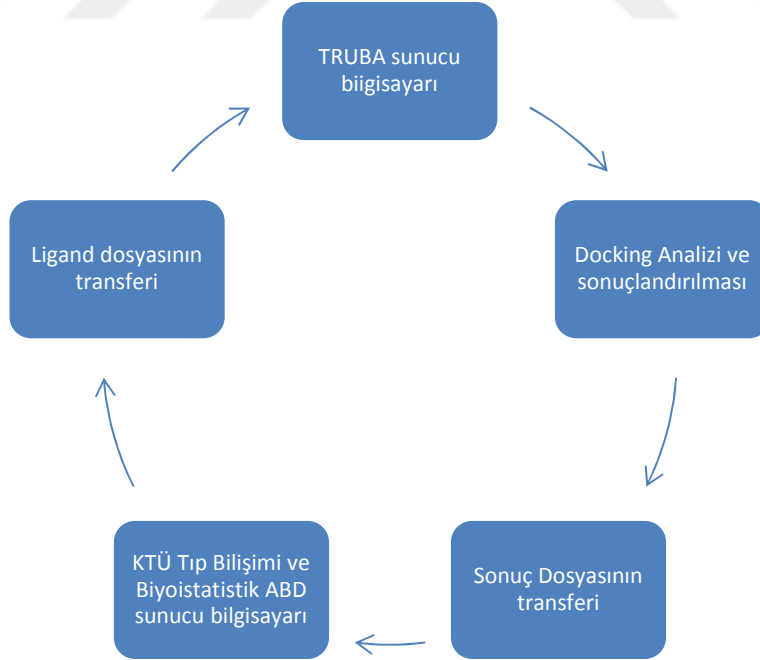
Resim 62. Sanal tarama analizi için aday reseptör dosyalarından seçim yapılması ve seçilen dosya içeriğinin analiz için hazır hale getirilmesi süreci. (a) “*ligands_list*” dosyası (b) “*receptor_preparation*” dosyası (c) “*box_params*” dosyası (d) Kod betiği çalışma sürecinden bir kesit. (e) “*vs_inputs*” dosya içeriğinden bir kesit. (f) “*local_run*” sonuç dosyalarından bir kesit. (g) “*web_run*” sonuç dosyalarından bir kesit.

00_01_97_28200207_vina_out.pdbqt										00_01_97_28200207_vina_out.pdbqt										
999 GB Volume / media/burak/5276089-0210-4779-b77a-fc0b2022c3/Thesis_data/docking_part/log/vs_accuracy/web_run/ty26_protein.pdbqt										999 GB Volume / media/burak/5276089-0210-4779-b77a-fc0b2022c3/Thesis_data/docking_part/log/vs_accuracy/web_run/ty26_protein.pdbqt										
1	MODEL	1								1	MODEL	1								
2	REMARK	VINA RESULT: -6.8 0.000 0.000								2	REMARK	VINA RESULT: -5.8 0.000 0.000								
3	REMARK	Name = ZINC0000000944								3	REMARK	Name = ZINC0000000944								
4	REMARK	3 active torsions								4	REMARK	3 active torsions								
5	REMARK	status: 'A' for Active; 'I' for Inactive								5	REMARK	status: 'A' for Active; 'I' for Inactive								
6	REMARK	1 A between atoms: C1.1 and N2.2								6	REMARK	1 A between atoms: C1.1 and N2.2								
7	REMARK	2 A between atoms: N2.2 and C1.3								7	REMARK	2 A between atoms: N2.2 and C1.3								
8	REMARK	3 A between atoms: C4.4 and C6.6								8	REMARK	3 A between atoms: C4.4 and C6.6								
9	REMARK									9	REMARK									
10	REMARK									10	REMARK									
11	REMARK									11	REMARK									
12	ATOM	1 C6 ZIN A 1	34.318	-12.954	-1.200	0.00	0.00	-0.140	A	12	ATOM	1 C6 ZIN A 1	33.476	-17.262	0.628	0.00	0.00	-0.140	A	
13	ATOM	2 C2 ZIN A 1	33.719	-16.996	-0.370	0.00	0.00	-0.120	A	13	ATOM	2 C2 ZIN A 1	33.878	-17.858	-0.370	0.00	0.00	-0.120	A	
14	ATOM	3 C6 ZIN A 1	34.523	-18.761	-1.139	0.00	0.00	-0.130	A	14	ATOM	4 C6B ZIN A 1	34.759	-18.260	-0.200	0.00	0.00	-0.110	A	
15	ATOM	4 C6B ZIN A 1	34.759	-18.260	-0.200	0.00	0.00	-0.110	A	15	ATOM	5 C6B ZIN A 1	34.759	-18.260	-0.200	0.00	0.00	-0.110	A	
16	ATOM	5 C1B ZIN A 1	33.689	-19.474	-0.923	0.00	0.00	-0.120	A	16	ATOM	6 C11 ZIN A 1	33.683	-18.874	1.035	0.00	0.00	-0.120	A	
17	ATOM	6 C11 ZIN A 1	33.683	-18.874	1.035	0.00	0.00	-0.120	A	17	ATOM	7 C11 ZIN A 1	33.683	-18.874	1.035	0.00	0.00	-0.120	A	
18	ATOM	7 C12 ZIN A 1	33.683	-18.874	1.035	0.00	0.00	-0.120	A	18	ATOM	8 C12 ZIN A 1	33.683	-18.874	1.035	0.00	0.00	-0.120	A	
19	ATOM	8 C12 ZIN A 1	33.683	-18.874	1.035	0.00	0.00	-0.120	A	19	ATOM	9 C12 ZIN A 1	33.683	-18.874	1.035	0.00	0.00	-0.120	A	
20	ATOM	9 C13 ZIN A 1	33.683	-18.874	1.035	0.00	0.00	-0.120	A	20	ATOM	10 C13 ZIN A 1	33.683	-18.874	1.035	0.00	0.00	-0.120	A	
21	ATOM	10 C13 ZIN A 1	33.683	-18.874	1.035	0.00	0.00	-0.120	A	21	ATOM	11 C13 ZIN A 1	33.683	-18.874	1.035	0.00	0.00	-0.120	A	
22	ENDMODEL									22	ENDMODEL									
23	BRANCH	1 14								23	BRANCH	1 14								
24	ATOM	11 C1 ZIN A 1	33.783	-21.303	-1.476	0.00	0.00	-0.200	A	24	ATOM	11 C1 ZIN A 1	33.878	-18.762	0.600	0.00	0.00	-0.200	A	
25	ATOM	12 N2 ZIN A 1	33.783	-21.303	-1.476	0.00	0.00	-0.200	A	25	ATOM	12 N2 ZIN A 1	33.812	-18.867	0.803	0.00	0.00	-0.200	A	
26	ATOM	13 C1 ZIN A 1	33.783	-21.303	-1.476	0.00	0.00	-0.200	A	26	ATOM	13 C1 ZIN A 1	33.812	-18.867	0.803	0.00	0.00	-0.200	A	
27	ATOM	14 C4 ZIN A 1	33.783	-21.303	-1.476	0.00	0.00	-0.200	A	27	ATOM	14 C4 ZIN A 1	33.744	-18.138	0.677	0.00	0.00	-0.200	A	
28	ENDBRANCH									28	ENDBRANCH	1 14	33.753	-16.325	0.466	0.00	0.00	-0.200	A	
29	ENDBRANCH	1 14								29	ENDBRANCH	1 14								
30	TORSIONSET	1 14								30	TORSIONSET	1 14								
31	ENDTORS									31	ENDTORS									
32	MODEL	2								32	MODEL	2								
33	REMARK	VINA RESULT: -5.9 2.000 3.139								33	REMARK	VINA RESULT: -5.8 4.856 6.761								
34	REMARK	Name = ZINC0000000944								34	REMARK	Name = ZINC0000000944								
35	REMARK	3 active torsions								35	REMARK	3 active torsions								
36	REMARK	status: 'A' for Active; 'I' for Inactive								36	REMARK	status: 'A' for Active; 'I' for Inactive								
37	REMARK	1 A between atoms: C1.1 and N2.2								37	REMARK	1 A between atoms: C1.1 and N2.2								
38	REMARK	2 A between atoms: N2.2 and C1.3								38	REMARK	2 A between atoms: N2.2 and C1.3								
39	REMARK	3 A between atoms: C4.4 and C6.6								39	REMARK	3 A between atoms: C4.4 and C6.6								
40	REMARK									40	REMARK									
41	REMARK									41	REMARK									
42	ATOM	1 C6 ZIN A 1	34.242	-20.252	-1.334	0.00	0.00	-0.140	A	42	ATOM	1 C6 ZIN A 1	33.684	-20.094	-1.337	0.00	0.00	-0.140	A	
43	ATOM	2 C2 ZIN A 1								43	ATOM	2 C2 ZIN A 1	32.552	-20.944	-1.346	0.00	0.00	-0.120	A	
b										b										
a										a										
c										c										

çalışmamızı olumsuz yönde direkt etkilemiş ve yöntemimizde yeni bir düzenleme yapma gerekliliği doğurmuştur.

İlk planda kullanılan kod betiği tüm ilaç adayı moleküllerin TRUBA kaynağında bize tahsis edilmiş disk alanına yüklenmesi ve ardından sırasıyla her bir moleküle ait sonuç dosyasının elde edilmesi şeklindeydi. Sonuç dosyaları da analizler tamamlanıncaya kadar TRUBA kaynağında bize tahsis edilen alanda saklanacak ve analizler tamamlandıktan sonra tüm dosyaların yerel diske aktarımı sağlanacaktı. Ancak yukarıda bahsedilen kota kısıtlaması nedeniyle kod betiğinin çalışma düzeninde değişiklik yapılmıştır. Yeni süreçte yapılan işlemler aşağıda açıklanmıştır.

Tüm ilaç adayı molekül dosyaları KTÜ Tıp Bilişimi ve Biyoistatistik ABD sunucu bilgisayarına aktarılmıştır. Ayrıca sanal tarama analizi için gerekli girdi dosyaları da yine aynı kaynağa aktarılmıştır. Bu dosyalar sanal tarama analizi için önce TRUBA sunucu bilgisayarına transfer edilmiş, ardından analiz yapılmıştır. Sanal tarama analizinden elde edilen sonuç dosyası da tekrar KTÜ Tıp Bilişimi ve Biyoistatistik ABD sunucu bilgisayarındaki ilgili konumuna transfer edilmiştir.

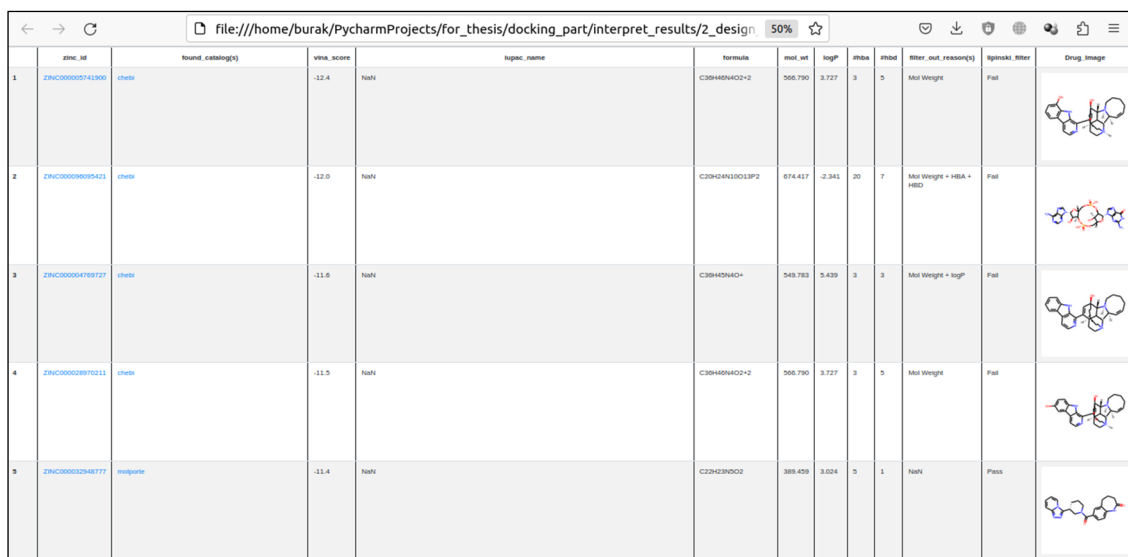


Şekil 23. Dosya sayı kısıtlaması nedeniyle geliştirilen yeni yöntemle TRUBA sunucu kaynağı kullanılarak sanal tarama analizinin yapılma süreci.

3.5.5. Sanal Tarama Analizi Sonuç Dosyalarının Raporlanması

Sanal tarama analizleri “vina”(76, 77) programı ile yapılmıştır. AD4 yerine “vina”(76, 77) tercih edilme nedeni çalışma hızı ve performansı açısından daha iyi sonuçlar elde etmiş olduğunun üreticileri tarafından belirtilmesidir. Bahsi geçen büyük performans farkının, “vina”(76, 77) yazılımında kullanılan BFGS gradyan tabanlı optimize edici ile birleştirilmiş bir Monte-Carlo (MC) yinelemeli arama algoritmasından kaynaklandığı ifade edilmiştir (77).

Elde edilen sonuç dosyaları her bir ilaç adayı molekülün belirlediğimiz reseptör dosyasına en iyi şekilde bağlanabildiği bölgenin konumunu ve bağlanma enerjisini içermektedir. Tüm dosyalarda benzer içerikler yer almaktadır. Bu kısım için hazırlanan kod betiği her bir sonuç dosyası içerisinde gerekli kısımları okumaktadır. Ayrıca okumuş olduğu ilaç adayı molekül dosyasına ait bazı kimyasal hesaplamaları ve özellikleri de elde etmektedir. Bu işlemler için “Rdkit” paketi kullanılmıştır. Elde edilen hesaplamalı kimyasal veriler “Lipinski’nin Beş Kuralı (RO5)” içinde yer alan hesaplamaları da kapsamaktadır. Yine “Rdkit” paketi kullanılarak çalışmada kullanılan ilaç adayı moleküllerin iki boyutlu kimyasal çizimleri de elde edilmiştir. Tüm bu veriler “csv” dosya düzeninde kaydedilmiş ve ayrıca “html” dosya düzeninde raporlanmıştır. Dosya içeriği oldukça fazla kayıt içerdiği için html rapor çıktısının araştırmacılar tarafından incelenmesi önerilir (Bkz. Ek 1). Resim 65’te elde edilen sonuçların temsili bir kısmı verilmiştir.



	zinc_id	source_catalog(s)	vina_score	ligand_name	formula	mol_wt	logp	hba	hbd	filter_out_reason(s)	lipinski_filter	Drug_Image
1	ZINC00000741800	chelix	-12.4	NaN	C38H48N4O2+2	566.790	3.727	3	2	Mol Weight	Fail	
2	ZINC00000909421	chelix	-12.0	NaN	C39H48N4O213P2	674.417	-2.341	20	7	Mol Weight + HBA + HBD	Fail	
3	ZINC000004798727	chelix	-11.6	NaN	C38H48N4O2+	549.783	5.439	3	3	Mol Weight + logP	Fail	
4	ZINC00002897011	chelix	-11.5	NaN	C38H48N4O2+2	566.790	3.727	3	3	Mol Weight	Fail	
5	ZINC00002948777	molport	-11.4	NaN	C22H28N2O2	388.458	3.024	5	1	NaN	Pass	

Resim 65. Sanal tarama analizi sonuçları html rapor dosyası içeriğinden bir kesit.

4. BULGULAR

4.1. Riboşalter Aile Üyesi RNA Elemanların Taksonomik Dağılımları

Bu kısımda yer alan veri seti 35 291 adet genom sekans verisinden oluşmaktadır. Riboşalter aile üyesi 40 adet RNA elemanlarının büyük oranla bakteri türü genomlarda yer aldığı Tablo 11 incelendiğinde kolaylıkla görülebilir.

Tablo 11. Riboşalterlerin kaynak genomlara dağılımı

Rfam Kimliği	Bakteri n (%)	Ökaryot n (%)	Arkea n (%)	Virus n (%)	Diğer n (%)	Toplam n (%)
RF01056	25 (89.29)	3 (10.71)	0 (0)	0 (0)	0 (0)	28 (100)
RF00379	1 238 (98.72)	2 (0.16)	0 (0)	0 (0)	14 (1.12)	1 254 (100)
RF00059	5 473 (91.55)	440 (7.36)	30 (0.5)	0 (0)	35 (0.59)	5 978 (100)
RF00168	1 308 (98.87)	9 (0.68)	0 (0)	0 (0)	6 (0.45)	1 323 (100)
RF01750	1 604 (99.32)	2 (0.12)	0 (0)	0 (0)	9 (0.56)	1 615 (100)
RF01725	296 (98.67)	4 (1.33)	0 (0)	0 (0)	0 (0)	300 (100)
RF00521	607 (99.67)	2 (0.33)	0 (0)	0 (0)	0 (0)	609 (100)
RF01786	259 (99.62)	0 (0)	0 (0)	0 (0)	1 (0.38)	260 (100)
RF00080	683 (97.85)	9 (1.29)	0 (0)	0 (0)	6 (0.86)	698 (100)
RF00504	2 905 (98.91)	14 (0.48)	1 (0.03)	0 (0)	17 (0.58)	2 937 (100)
RF00167	909 (98.38)	3 (0.32)	0 (0)	0 (0)	12 (1.3)	924 (100)
RF01767	179 (98.9)	0 (0)	0 (0)	0 (0)	2 (1.1)	181 (100)
RF01689	188 (97.41)	1 (0.52)	1 (0.52)	0 (0)	3 (1.55)	193 (100)
RF01727	199 (98.03)	2 (0.99)	0 (0)	0 (0)	2 (0.99)	203 (100)
RF00634	631 (98.75)	0 (0)	0 (0)	0 (0)	8 (1.25)	639 (100)
RF00050	2 827 (98.95)	12 (0.42)	8 (0.28)	0 (0)	10 (0.35)	2 857 (100)
RF00162	2 422 (98.58)	11 (0.45)	0 (0)	0 (0)	24 (0.98)	2 457 (100)
RF01510	1 (100)	0 (0)	0 (0)	0 (0)	0 (0)	1 (100)
RF00380	652 (97.9)	3 (0.45)	1 (0.15)	0 (0)	10 (1.5)	666 (100)
RF01831	504 (99.41)	0 (0)	0 (0)	0 (0)	3 (0.59)	507 (100)
RF00442	646 (99.23)	3 (0.46)	0 (0)	0 (0)	2 (0.31)	651 (100)
RF01054	45 (93.75)	1 (2.08)	0 (0)	0 (0)	2 (4.17)	48 (100)
RF01826	8 (72.73)	1 (9.09)	0 (0)	0 (0)	2 (18.18)	11 (100)
RF01482	328 (99.09)	1 (0.3)	0 (0)	0 (0)	2 (0.6)	331 (100)
RF01739	119 (92.97)	0 (0)	0 (0)	9 (7.03)	0 (0)	128 (100)
RF01055	655 (99.24)	5 (0.76)	0 (0)	0 (0)	0 (0)	660 (100)
RF00174	5 009 (98.84)	29 (0.57)	4 (0.08)	0 (0)	26 (0.51)	5 068 (100)
RF01057	646 (98.33)	6 (0.91)	0 (0)	0 (0)	5 (0.76)	657 (100)
RF00522	391 (98.99)	1 (0.25)	0 (0)	0 (0)	3 (0.76)	395 (100)
RF00234	885 (99.55)	0 (0)	0 (0)	0 (0)	4 (0.45)	889 (100)
RF01734	1 525 (94.08)	16 (0.99)	76 (4.69)	0 (0)	4 (0.25)	1 621 (100)
RF01704	62 (81.58)	1 (1.32)	0 (0)	13 (17.11)	0 (0)	76 (100)
RF02680	6 (85.71)	0 (0)	0 (0)	0 (0)	1 (14.29)	7 (100)
RF02683	150 (97.4)	0 (0)	0 (0)	0 (0)	4 (2.6)	154 (100)

Tablo 11. (Devam)

RF02885	13 (76.47)	0 (0)	0 (0)	0 (0)	4 (23.53)	17 (100)
RF02912	20 (95.24)	1 (4.76)	0 (0)	0 (0)	0 (0)	21 (100)
RF03057	271 (97.83)	4 (1.44)	0 (0)	0 (0)	2 (0.72)	277 (100)
RF03058	252 (98.82)	3 (1.18)	0 (0)	0 (0)	0 (0)	255 (100)
RF03071	116 (98.31)	0 (0)	0 (0)	0 (0)	2 (1.69)	118 (100)
RF03072	0 (0)	0 (0)	0 (0)	0 (0)	277 (100)	277 (100)
Ribo Şalter	34 057 (96.5)	589 (1.67)	121 (0.34)	22 (0.06)	502 (1.42)	35 291 (100)

4.2. GC-Oranı Analizi İçin Elde Edilen İstatistiksel Sonuç Dağılımları

Bu kısımda riboşalter sekans dizilimindeki GC-oran değeri ile kaynak genomuna ait sekans dizilimindeki GC-oranı karşılaştırmalarına ait analiz bulguları verilmiştir. Bulgular incelendiğinde iki oran arasında fark olmadığını savunan H_0 hipotezinin genellikle reddedildiği bulgular elde edildiği görülmüştür (Tablo 12).

Tablo 12. Tek örneklem t-testi p istatistiğine göre alınan karar dağılımları

<i>Rfam</i> kimliği	H_0 Kabul n (%)	H_0 Ret n (%)	Toplam n (%)	<i>Rfam</i> kimliği	H_0 Kabul n (%)	H_0 Ret n (%)	Toplam n (%)
RF00050	414 (10.1)	3683 (89.9)	4097 (100)	RF01510	0 (0)	4 (100)	4 (100)
RF00059	1557 (15.38)	8566 (84.62)	10123 (100)	RF01689	16 (11.76)	120 (88.24)	136 (100)
RF00080	193 (24.13)	607 (75.88)	800 (100)	RF01704	10 (14.93)	57 (85.07)	67 (100)
RF00162	594 (10.04)	5325 (89.96)	5919 (100)	RF01725	71 (20.46)	276 (79.54)	347 (100)
RF00167	474 (18.57)	2079 (81.43)	2553 (100)	RF01727	23 (11.68)	174 (88.32)	197 (100)
RF00168	258 (11.42)	2001 (88.58)	2259 (100)	RF01734	361 (19.24)	1515 (80.76)	1876 (100)
RF00174	2673 (19.38)	11120 (80.62)	13793 (100)	RF01739	25 (18.25)	112 (81.75)	137 (100)
RF00234	79 (8.4)	862 (91.6)	941 (100)	RF01750	297 (18.38)	1319 (81.62)	1616 (100)
RF00379	640 (16.42)	3258 (83.58)	3898 (100)	RF01767	32 (18.6)	140 (81.4)	172 (100)
RF00380	104 (10.71)	867 (89.29)	971 (100)	RF01786	63 (15.14)	353 (84.86)	416 (100)
RF00442	116 (13.29)	757 (86.71)	873 (100)	RF01826	1 (8.33)	11 (91.67)	12 (100)
RF00504	770 (16.91)	3783 (83.09)	4553 (100)	RF01831	49 (8.84)	505 (91.16)	554 (100)
RF00521	69 (8.22)	770 (91.78)	839 (100)	RF02680	0 (0)	9 (100)	9 (100)
RF00522	88 (23.47)	287 (76.53)	375 (100)	RF02683	13 (7.65)	157 (92.35)	170 (100)
RF00634	248 (20.31)	973 (79.69)	1221 (100)	RF02885	0 (0)	8 (100)	8 (100)
RF01054	15 (28.3)	38 (71.7)	53 (100)	RF02912	8 (28.57)	20 (71.43)	28 (100)
RF01055	257 (23.24)	849 (76.76)	1106 (100)	RF03057	28 (10.18)	247 (89.82)	275 (100)
RF01056	12 (29.27)	29 (70.73)	41 (100)	RF03058	42 (14.58)	246 (85.42)	288 (100)
RF01057	178 (21.07)	667 (78.93)	845 (100)	RF03071	8 (6.67)	112 (93.33)	120 (100)
RF01482	32 (19.05)	136 (80.95)	168 (100)	RF03072	19 (7.72)	227 (92.28)	246 (100)

4.3. Reseptör Adayı Riboşaltere Ait Üç Boyutlu Yapı Dosyalarının Taksonomik Dağılımı

Bu kısımda Rfam veri tabanında belirtilen riboşalterler ile bağlantılı üç boyutlu yapı dosyalarının kaynak genomun “*superkingdom*” taksonomi seviyesine göre dağılımları verilmiştir (Tablo 13).

Tablo 13. Reseptör adayı riboşalterlere ait üç boyutlu yapı dosyalarının taksonomik dağılımı

<i>Rfam</i> kimliği	Bakteri n (%)	Ökaryot n (%)	Diğer n (%)	Toplam n (%)
RF00059	0 (0)	4 (80)	1 (20)	5 (100)
RF00162	1 (50)	0 (0)	1 (50)	2 (100)
RF00167	3 (100)	0 (0)	0 (0)	3 (100)
RF00168	2 (100)	0 (0)	0 (0)	2 (100)
RF00174	1 (100)	0 (0)	0 (0)	1 (100)
RF00379	4 (26.67)	11 (73.33)	0 (0)	15 (100)
RF00504	1 (50)	1 (50)	0 (0)	2 (100)
RF01727	1 (100)	0 (0)	0 (0)	1 (100)
RF01831	1 (100)	0 (0)	0 (0)	1 (100)
RF02680	1 (100)	0 (0)	0 (0)	1 (100)
Toplam	15 (45.45)	16 (48.48)	2 (6.06)	33 (100)

4.4. RNA Central Biyolojik Veri Tabanı Üzerinden “Anotasyon” Analizi Süreci Sonunda Elde Edilen Bulgular

Bu kısımda elde edilen bulguların tamamı eklerde belirtilen dijital bağlantılarda verilmiştir (Bkz. Ek 1).

Tablo 14. Süreç sonunda elde edilen anlamlı kabul edilebilecek sınırdaki bulgular

Sorgu No	Eşleşen Bölge	Sorgu Kaynağı	Bit Skoru	E-değeri
1	Homo sapiens contig freeze2_11051 genomic sequence (GenBank: GU266878.1)	Lysine riboswitch (RF00168)	109.0	1.3e ⁻²⁹
2	Homo sapiens contig freeze2_11051 genomic sequence (GenBank: GU266878.1)	Lysine riboswitch (RF00168)	109.0	1.3e ⁻²⁹
3	Homo sapiens contig freeze2_11051 genomic sequence (GenBank: GU266878.1)	Lysine riboswitch (RF00168)	46.6	1.5e ⁻¹¹

Tablo 15. Tablo 14’te verilen BLAST sorgu numaralarının açıklaması

Sorgu No	Sorgu Açıklaması
1	34 163-RfamId: RF00168, SeqAcc: AEYM01001934.1, Specie is: Lactobacillus rhamnosus MTCC 5 462 contig45, whole genome shotgun sequence., From: 1 053 - To: 874
2	33 319-RfamId: RF00168, SeqAcc: CP000423.1, Specie is: Lactobacillus casei ATCC 334, complete genome., From: 92 631 - To: 92 453
3	32 331-RfamId: RF00168, SeqAcc: CAGH01000008.1, Specie is: Weissella confusa LBAE C39-2, WGS project CAGH01000000 data, contig: WeissC39_1_contig00008, From: 71 708 - To: 71 539

4.5. ZINC Veri Tabanından Çekilen İlaç Adayı Molekül Dosyalarına Ait Çeşitli Özellik Dağılımları

Bu kısımda çalışmada kullanılan ilaç adayı moleküllere ait çeşitli özellik dağılım sıklıkları verilmiştir (Tablo 16-21).

Tablo 16. İlaç adayı molekül dosyalarının moleküler ağırlığına göre dağılımı

Değişkenin		Sıklık	
Kodu	Değeri	n	%
A	<200	185 912	9.77
B	200 - 250	214 595	11.28
C	250 - 300	367 633	19.32
D	300 - 325	281 275	14.78
E	325 ve 350	339 215	17.83
F	350 ve 375	287 718	15.12
G	375 ve 400	220 554	11.59
H	400 ve 425	1 059	0.06
I	425 ve 450	751	0.04
J	450 ve 500	1 264	0.07
K	>500	2 493	0.13
Toplam		1 902 469	100

Tablo 17. İlaç adayı molekül dosyalarının *logP* (*xlogP*) değerine göre dağılımı

Değişkenin		Sıklık	
Kodu	Değeri	n	%
A	<-1	42 211	2.22
B	-1 - 0	84 985	4.47
C	0 - 1	245 177	12.89
D	1 - 2	516 674	27.16
E	2 - 2.5	352 988	18.55
F	2.5 - 3	354 386	18.63
G	3 - 3.5	299 440	15.74
H	3.5 - 4	1 304	0.07
I	4 - 4.5	1 258	0.07
J	4.5 - 5	1 023	0.05
K	>5	3 023	0.16
Toplam		1 902 469	100

Tablo 18. İlaç adayı molekül dosyalarının *reactivity* değişkenine göre dağılımı

Değişkenin		Sıklık	
Kodu	Değeri	n	%
A	<i>Anodyne</i>	1 645 637	86.5
B	<i>Bother</i>	74 293	3.91
C	<i>Clean</i>	32 764	1.72
E	<i>Mild Reactivity</i>	129 043	6.78
G	<i>Reactive</i>	11 794	0.62
I	<i>Hot Chemistry</i>	8 938	0.47
Toplam		1 902 469	100

Tablo 19. İlaç adayı molekül dosyalarının *purchasability* değişkenine göre dağılımı

Değişkenin		Sıklık	
Kodu	Değeri	n	%
A	<i>In Stock</i>	1 252 017	68.95
B	<i>In Stock</i>	83 341	4.59
C	<i>Agent</i>	2 701	0.15
D	<i>On Demand</i>	408 644	22.5
E	<i>Boutique</i>	65 152	3.59
F	<i>Not For Sale</i>	4 046	0.22
Toplam		1 815 901	100

Tablo 20. İlaç adayı molekül dosyalarının *pH seviyesi* değişkenine göre dağılımı

Değişkenin		Sıklık	
Kodu	Değeri	n	%
H	High, (pH 8.4 - 9.4)	127 494	6.7
R	Ref, (pH=7.4)	1 503 746	79.0
M	Mid, (pH 6.4 - 8.4)	176 571	9.28
L	Low, (pH 5.4 - 6.4)	94 658	4.98
Toplam		1 902 469	100

Tablo 21. İlaç adayı molekül dosyalarının *net moleküler yük* değişkenine göre dağılımı

Değişkenin		Sıklık	
Kodu	Değeri	n	%
M	=-1	659 022	34.64
L	<=-2	45 673	2.4
N	=0	959 099	50.41
O	=+1	219 862	11.56
P	>=+2	18 813	0.99
Toplam		1 902 469	100

4.6. ZINC Veri Tabanından Çekilen İlaç Adayı Molekül Dosyalarının “*Lipinski’nin Beş Kuralına*” Göre Dağılımı

Bir molekülün ilaç adayı kabul edilmesi için uyması gereken literatürde belirlenmiş bazı kuralları vardır. Bu kısımda ilk olarak çalışmada kullanılan ilaç adayı moleküllerin *Lipinski’nin Beş Kuralı’na (RO5)* uyum dağılımı verilmiştir (83, 84). Ardından literatürden derlediğimiz FDA onayı almış ve aktif olarak kullanılan çeşitli ilaçların RO5 uyum dağılımı verilmiştir (83, 84). Tablo verilen bulgular tartışma ve sonuç kısmında tartışılmıştır.

Tablo 22. Çalışmada kullanılan molekül dosyalarının RO5 uyumuna göre dağılımı

Dosya Türü	Kurala Uyan	Kurala Uymayan	Toplam
FDA onaylı ilaçlar	2 047 (%68.53)	940 (%31.47)	2 987 (%100)
Sanal tarama analizinde kullanılanlar	1 320 924 (%95.68)	5891 (%4.32)	1 362 815 (%100)

4.7. *V. vulnificus* Kaynaklı Hastalık Tedavisinde Kullanılan İlaçlar ve Diğer Moleküllerle Olan *Tanimoto* Benzerlikleri

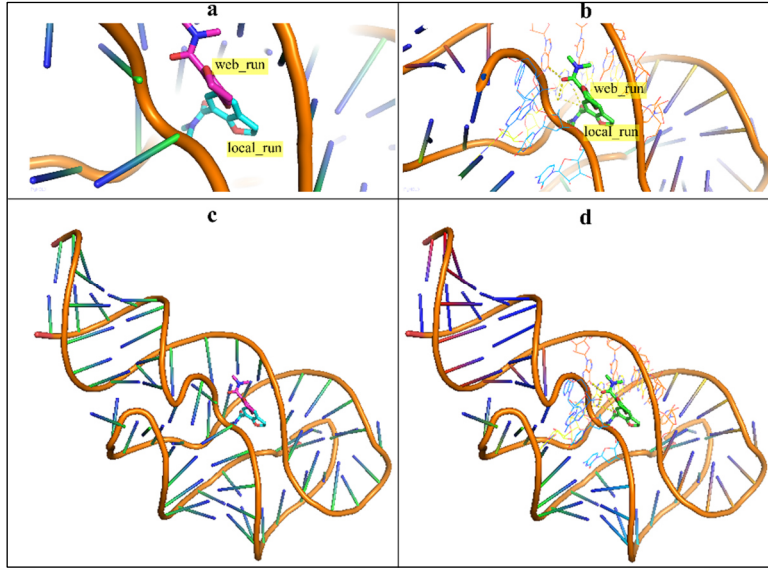
Tablo 23. *V. vulnificus* tedavisinde kullanılan ilaçlar ve bu ilaçlara benzer moleküllerin *Tanimoto* Benzerliği Tanımlayıcı İstatistikleri (n=1000)

İlaç Adı	Min	Ort ± Ss	Mak
<i>Doxycycline</i>	0.758	0.943±0.031	1.000
<i>Vibramycin</i>	0.758	0.943±0.031	1.000
<i>Ceftazidime</i>	0.618	0.886±0.073	1.000
<i>Ciprofloxacin</i>	0.407	0.831±0.100	1.000
<i>Cefotaxime</i>	0.701	0.915±0.049	1.000

4.8. Sanal Tarama Analizlerinin Doğruluğunun Belirlenmesi

Bu kısımda elde edilen bulgular için sınıf içi korelasyon katsayısı kullanıldı. İki değerlendiricinin aynı olayı değerlendirdiği bir çalışma düzenine uygun olacak şekilde tasarlanan veriler, daha önceden kod betiği ile elde edilmişti. Çalışma düzeni, değerlendiricilerin önceden belli olduğu deneklerin ise rastgele seçildiği “*Two-Way Mixed*” olarak belirlendi. İki sanal tarama analizi sonuçlarının elde ettiği bulgularda tam bir uyum olup olmadığını araştırdığımız için, çalışma türü “*Absolute Agreement*” olarak belirlendi.

Veri seti içerisinde ardışık düzen ile yazar tarafından elde edilen sanal tarama analizi sonuç verileri “*local*” ön adlı; “*webina*”(117) moleküler yansıtırma sunucusu ile elde edilen sanal tarama analizi sonuç verileri “*web*” ön adlı değişkenler ile temsil edildi. Değişkenlerde tutulan veriler ise ilaç adayı moleküllerin reseptör üzerinde üç boyutlu koordinat eksenindeki en iyi moleküler yansımalar pozisyonlarına karşılık gelmektedir. Bu çalışma için gerekli örneklem sayısı ilgili literatür kaynağında yer alan kriterler dikkate alınarak belirlenmiştir (119). Bu kısımda elde edilen bulgular için Tablo 24 ve Resim 66 incelenebilir.



Resim 66. Sınıf içi korelasyon analizi verilerinden örnek gösterimler. **(a)** Örnek sonuç dosyasının iki analiz sonucundaki konumları yakın çekim. **(b)** Örnek sonuç dosyasının iki analiz sonucundaki farklı konformasyonları (ligand etkileşimleri) yakın çekim. **(c)** Örnek sonuç dosyasının iki analiz sonucundaki konumları uzak çekim. **(d)** Örnek sonuç dosyasının iki analiz sonucundaki farklı konformasyonları (ligand etkileşimleri) uzak çekim.

Tablo 24. Sanal tarama analiz sonuçlarına dair sınıf içi korelasyon istatistikleri

Ölçüm türü	Sınıf içi Korelasyon Katsayısı	%95 Güven Aralığı		Gerçek Değerin 0 olduğuna dair F Testi			
	r	Alt Sınır	Üst Sınır	F değeri	sd1	sd2	p
X eksen	0.465	0.419	0.507	1.905	3 311	3 311	<0.001
Y eksen	0.592	0.564	0.619	2.454	3 311	3 311	<0.001
Z eksen	0.497	0.462	0.531	1.997	3 311	3 311	<0.001

*sd, serbestlik derecesini ifade eder

Tablo 24’te elde edilen sınıf içi güvenilirlik katsayıları *Cichetti* (120) ölçeğine göre yorumlandığında üç eksen de yeterli düzeyde uyum olduğu gözlenmiştir. Bu beklenen bir bulgudur. Daha yüksek seviyede bir uyum elde edilememesinin nedeniyse, kullandığımız üç boyutlu moleküler yanaşma verilerinin kendine has durumudur. Bu durum kullanılan moleküler yanaştırma yazılımının kendi web sayfasında yer alan sıkça sorulanlar kısmında şöyle ifade edilmiştir:

“Yanaştırma algoritması deterministik değildir. Bu reseptör-ligand çifti ile puanlama fonksiyonunun minimumu doğru konformasyona karşılık gelse de yanaştırma algoritması bazen onu bulmakta başarısız olur. Birkaç

kez deneyin ve kendiniz görün. Minimumu bulamama olasılığının farklı bir sistemle farklı olabileceğini unutmayın (121).”

4.9. Lipinski'nin Beş Kuralı Dikkate Alındığında En İyi Bağlanma Enerjisine Sahip İlk On Molekül

Tablo 25. Sanal tarama analizi sonucunda elde edilen en iyi bağlanma enerjisi'ne sahip ilk on molekül (*Lipinski'nin beş kuralı dikkate alınmış*)

ZINC Erişim Kimliği	Bulunduğu Kataloglar	Formülü	Vina Skoru
ZINC000032948777	Molporte	C ₂₂ H ₂₃ N ₅ O ₂	-11.4
ZINC000261361408	Specs	C ₁₄ H ₈ N ₁₀ O ₄ ⁻²	-11.1
ZINC000000956909	Iflab	C ₁₈ H ¹² N ₅ O ₅ ⁺	-10.8
ZINC000952868194	Chbrp	C ₂₄ H ₂₆ N ₇ O ₅ ⁺	-10.7
ZINC000040394717	Chebi	C ₂₃ H ₂₃ NO ₇ S	-10.7
ZINC000056871308	Chebi	C ₃₀ H ₅₀ N ₃ O ⁺	-10.6
ZINC000001153249	Molporte	C ₁₈ H ₁₁ N ₃ O ₆	-10.5
ZINC000004262791	Molmall,Specs	C ₂₁ H ₁₇ O ₈ ⁻	-10.5
ZINC000004006841	Chbre,Specs,Molporte	C ₁₄ H ₈ N ₁₀ O ₄ ⁻²	-10.5
ZINC000004097280	Chebi	C ₂₂ H ₁₉ O ₁₀ ⁻	-10.4

Tablo 26. En iyi bağlanma enerjisi'ne sahip ilk on molekül ile *V. vulnificus* hastalığı tedavisinde kullanılan ilaç molekülleri arasındaki hesaplanmış *Tanimoto* benzerlikleri (*Lipinski'nin beş kuralı dikkate alınmış*)

ZINC Erişim Kimliği	Doxycycline	Vibramycin	Ceftazidime	Ciprofloxacin	Cefotaxime
ZINC000032948777	0.488	0.488	0.475	0.404	0.464
ZINC000261361408	0.338	0.338	0.3	0.267	0.285
ZINC000000956909	0.475	0.475	0.474	0.389	0.458
ZINC000952868194	0.589	0.589	0.581	0.444	0.547
ZINC000040394717	0.707	0.707	0.642	0.467	0.622
ZINC000056871308	0.266	0.266	0.241	0.206	0.232
ZINC000001153249	0.444	0.444	0.426	0.371	0.418
ZINC000004262791	0.654	0.654	0.58	0.457	0.559
ZINC000004006841	0.338	0.338	0.3	0.267	0.285
ZINC000004097280	0.71	0.71	0.638	0.494	0.611

4.10. Lipinski'nin Beş Kuralı Dikkate Alınmadığında En İyi Bağlanma Enerjisine Sahip İlk On Molekül

Tablo 27. Sanal tarama analizi sonucunda elde edilen en iyi bağlanma enerjisi'ne sahip ilk on molekül (*Lipinski'nin beş kuralı dikkate alınmamış*)

ZINC Erişim Kimliği	Bulunduğu Kataloglar	Formülü	Vina Skoru
ZINC000005741900	Chebi	$C_{36}H_{46}N_4O_2^{+2}$	-12.4
ZINC000096095421	Chebi	$C_{20}H_{24}N_{10}O_{13}P_2$	-12.0
ZINC000004769727	Chebi	$C_{36}H_{45}N_4O^+$	-11.6
ZINC000028970211	Chebi	$C_{36}H_{46}N_4O_2^{+2}$	-11.5
ZINC000032948777	Molporte	$C_{22}H_{23}N_5O_2$	-11.4
ZINC000085548520	Chebi	C_{60}	-11.3
ZINC000150340034	Chebi	$C_{14}F_{24}$	-11.2
ZINC000163930500	Chebi	$C_{14}F_{24}$	-11.2
ZINC000044352664	Chebi	$C_{31}H_{30}O_{10}$	-11.2
ZINC000084605049	Chebi	$C_{31}H_{39}N_8O_3^+$	-11.1

Tablo 28. En iyi bağlanma enerjisi'ne sahip ilk on molekül ile *V. vulnificus* hastalığı tedavisinde kullanılan ilaç molekülleri arasındaki hesaplanmış Tanimoto benzerlikleri (*Lipinski'nin beş kuralı dikkate alınmamış*)

ZINC Erişim Kimliği	Doxycycline	Vibramycin	Ceftazidime	Ciprofloxacin	Cefotaxime
ZINC000005741900	0.661	0.661	0.59	0.478	0.567
ZINC000096095421	0.595	0.595	0.569	0.447	0.549
ZINC000004769727	0.643	0.643	0.576	0.466	0.553
ZINC000028970211	0.653	0.653	0.584	0.469	0.561
ZINC000032948777	0.488	0.488	0.475	0.404	0.464
ZINC000085548520	0.036	0.036	0.033	0.05	0.033
ZINC000150340034	0.171	0.171	0.171	0.167	0.176
ZINC000163930500	0.171	0.171	0.171	0.167	0.176
ZINC000044352664	0.633	0.633	0.573	0.446	0.556
ZINC000084605049	0.589	0.589	0.558	0.447	0.537

5. TARTIŞMA ve SONUÇ

Bu çalışmanın sanal tarama analizi için kullanılan reseptör dosyasının ait olduğu *V. vulnificus* organizmasının insanlarda meydana getirdiği hastalıkların tedavisinde kullanılan beş farklı ilaç olduğu bilgisine literatürden ulaşılmıştır (122). İnsanda meydana gelen çeşitli hastalıkların tedavisinde kullanılan FDA onaylı ilaçların listesine ve içeriğine de literatürden ulaşılmıştır (123). Sanal tarama analizleri için kullanılacak ilaç adayı molekül dosyalarının uyması gereken çeşitli kriterler literatürde belirtilmiştir. Bunlardan en çok bilineni ve yaygın olarak kullanılanlarından birisi de kısaltması RO5 olan *Lipinski'nin Beş Kuralı'dır* (83, 84). Ancak elde ettiğimiz FDA onaylı ilaç veri seti RO5 açısından incelendiğinde tüm moleküllerin kurala uymadığı ve buna rağmen onaylı ilaç olarak kullanıldıkları görülmüştür (Tablo 22). Bu noktadan hareketle çalışmamızda elde ettiğimiz sanal tarama analizi sonuçları içerisinde RO5 uyan ve uymayan ilk on molekül bulgularına yer verilmiştir. Mevcut bulgular, *V. vulnificus* organizmasının insanda meydana getirdiği hastalıkların tedavisinde kullanılan ilaçlar açısından da tartışılmıştır. Tablo 26 ve 28 içerisinde hastalık tedavisinde kullanılan ilaçlar ile çalışma sonucunda elde edilen en iyi bağlanma enerjisine sahip moleküllerin *Tanimoto* benzerlikleri verilmiştir (85). Ayrıca *V. vulnificus* kaynaklı hastalık tedavisinde kullanılan ilaçların kendileriyle benzer molekül listelerine *PubChem* veri tabanından ulaşılmıştır (124). Bu listeler kullanılarak her bir ilacın literatürde kendine benzer kabul edilen ilaçlar ile olan *Tanimoto* benzerliği (85) değerleri de hesaplanmış ve Tablo 23'te verilmiştir. Literatürde iki molekülün benzer kabul edilebilmesi için sıfır ile bir arasında değer alabilen *Tanimoto* benzerliği değerinin en az "0.85" olarak hesaplanmış olması gerekmektedir (85). Tablo 23 incelendiğinde *PubChem* (124) kaynaklı listelerden derlenen sonuçların bu alt sınır değerine uyduğu gözlenmiştir.

Tablo 25 ve 26 incelendiğinde RO5 dikkate alınarak elde edilen *ZINC000040394717* erişim kimliğine sahip molekülün *Ceftazidime*; *ZINC000004097280* erişim kimliğine sahip molekülün ise *Ciprofloxacin* ile benzer moleküller olduğu söylenebilir. Bu yargıya, her iki molekülün benzer kabul edildiği ilaç molekülüne ait Tablo 23'te verilen *Tanimoto* skoru tanımlayıcı istatistiklerindeki minimum değer üzerinde benzerlik skoru almış olmaları üzerine varılmıştır.

Tablo 27 ve 28 incelendiğinde RO5 dikkate alınmadan elde edilen *ZINC000005741900* erişim kimliğine sahip molekülün *Ciprofloxacin* ile benzer moleküller olduğu söylenebilir. Bu yargıya, bahsi geçen molekülün benzer kabul edildiği ilaç

molekülüne ait Tablo 23'te verilen *Tanimoto* benzerliği (85) değeri tanımlayıcı istatistiklerindeki minimum değerin üzerinde benzerlik skoru almış olması üzerine varılmıştır.

Lizin (Lys, K) proteinlerde bulunan 22 aminoasitten biridir. *Lizin* kalsiyum emilimi, kas proteinlerinin inşası, ameliyat sonrası ve spor yaralanmaları sonrası iyileşme sürecinde, vücut tarafından hormonların, antikorların ve enzimlerin sentezlenişinde önemli role sahiptir (125). *Lizin* riboşalteri ([RF00168](#)), amino asit *lizin* için hassas bir sensör görevi gören belirli haberci RNA'larda bulunan bir metabolit bağlayıcı RNA elementidir.

Glutarik asidemi tip 1 (GAI), vücudun lizin, hidroksilizin ve triptofan amino asitlerini tamamen parçalayamadığı kalıtsal bir hastalıktır (126). GAI, GCDH geni tarafından kodlanan glutaril-CoA dehidrojenaz (GCDH) enziminin eksikliğinden kaynaklanan otozomal resesif bir hastalıktır. Tedavi edilmeyen çoğu hasta, yüksek morbidite ve mortalite ile ilişkili bilateral striatal yaralanmaya bağlı olarak baskın distoni ile karmaşık bir hareket bozukluğu geliştirir (127–130). Bu, genellikle muhtemelen iyi huylu bir bulaşıcı hastalık tarafından tetiklenen akut bir ensefalopatik krizin ardından veya sinsice ortaya çıkabilir (130). Tersine, bazı kişiler ömür boyu semptomlardan muaf kalabilir. Bu kadar geniş fenotipik heterojenite iyi anlaşılmamıştır ve spesifik genotipler veya toksik metabolitlerin atılım düzeyi ile ilişkili değildir (131).

Bu çalışmanın amaçlarından birisi de insan genomu içerisinde riboşalter yapılarıyla uyumlu bölgeler olup olmadığının tespit edilmesiydi. Bu doğrultuda çalışmamızda uygulanan yöntem üzerinden *lizin* riboşalteri ile anlamlı seviyede kabul edilebilecek düzeyde eşleşen bir insan genom bölgesi tespit edilmiş ve bulguları Tablo 14'te verilmiştir (Bkz. 3.3.6.2. RNA Central Biyolojik Veri Tabanı Üzerinden “Anotasyon” Analizi Süreci). Ancak bu eşleşme Rfam veri tabanında kullanılan ve aile üyeliğine dahil edilme eşik değerine uygun yeterliğe sahip değildir (46). Yine de dikkate değer olabileceğini düşündüğümüz bu bulguyu tartışabilmek amacıyla yukarıdaki alıntı paragrafta yer alan bilgiler literatürden derlenmiştir.

Özellikle *Lizin* aminoasidiyle bağlantılı olan *GAI* hastalığının tüm semptomları bazı insanlarda ağır bir şekilde görülürken, bazı insanlardaysa ömür boyu hiçbir semptom

görülmediği literatür kaynak incelemesinden çıkardığımız bir bulgudur (131). Bu bulgunun işaret ettiği heterojenitenin kaynağının, ribozalterlerin kendilerine has biyolojik yapılarıyla benzerlik gösterebileceğini düşünmekteyiz. Ancak literatürde bu varsayımımızı destekleyecek herhangi bir çalışmaya ya da bulguya rastlanılmamıştır.

Bu çalışma kaynak olarak aldığı çalışmalarda uygulanan yöntemlere benzer mantıkla planlanarak tasarlanmıştır (132–134). Ancak kaynak çalışmalarda izlenen işlem adımları bireysel olarak her adımda kullanıcının çeşitli işlemleri kendisinin yapmasıyla ilerlemektedir. Benzer şekilde sanal tarama için kullanılan reseptör dosyası protein veri tabanından alınmamıştır. Bunun yerine RNA katlanmasını tahmin eden (*RNA folding*) çeşitli uygulamalar ile üretilen dosyalar kullanılmıştır. Bizim çalışmamız ise kullanıcı etkileşimine gerek kalmadan tüm sürecin önceden tasarlanarak planlandığı ve bu plana uygun olarak yazılmış olan kod betikleri ile çalıştığı bir ardışık düzen içerisinde ilerlemektedir. Bu ardışık düzen içerisinde kullanılan girdi verileri *Rfam* veri tabanından derlenmiş ve sanal tarama analizi için de protein veri tabanında yer alan *1Y26* erişim kimliğine sahip dosya kullanılmıştır. Kaynak çalışmalara göre daha kapsamlı ve kendi hiyerarşik düzeninde ilerleyen çalışmamız ayrıca kullanıcısına her süreç adımı sonunda inceleyip yorumlaması için *html* formatında çıktı dosyaları da üretmektedir.

Sonuç olarak bu çalışmada hedeflediğimiz amaçlara ulaşılmıştır. Ardışık düzen uygulamasının ana çerçevesi oturtulmuş olup benzer çalışmalar farklı *Rfam* aile üyeleri için de hazırlanabilir. Sonraki aşamada gerekli kaynağın sağlanabileceği bir proje desteği alınmasıyla birlikte, yazar tarafından oluşturulan kod betikleri üzerinden çalışan ardışık düzen uygulamamıza yönelik bir web uygulamasının da geliştirilmesi planlanmaktadır.

6. KAYNAKLAR

1. Gelfand MS, Mironov AA, Jomantas J, Kozlov YI, Perumov DA (1999). A conserved RNA structure element involved in the regulation of bacterial riboflavin synthesis genes. *Trends in Genetics* 15(11): 439-442.
2. Henkin TM (2000). Transcription termination control in bacteria. *Current Opinion in Microbiology* 3(2): 149-153.
3. Mandal M, Breaker RR (2004). Gene regulation by riboswitches. *Nature Reviews Molecular Cell Biology* 5: 451-463.
4. Nudler E, Mironov AS (2004). The riboswitch control of bacterial metabolism. *Trends in Biochemical Sciences* 29(1): 11-17.
5. Kalvari I, Nawrocki EP, Ontiveros-Palacios N, Argasinska J, Lamkiewicz K, Marz M, Griffiths-Jones S, Toffano-Nioche C, Gautheret D, Weinberg Z, Rivas E, Eddy SR, Finn RD, Bateman A, Petrov AI (2021). Rfam 14: Expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Research* 49(D1): 192-200.
6. Kalvari I, Nawrocki EP, Argasinska J, Quinones-Olvera N, Finn RD, Bateman A, Petrov AI (2018). Non-Coding RNA Analysis Using the Rfam Database. *Current Protocols in Bioinformatics* 62(1): e51.
7. Sherwood AV, Henkin TM (2016). Riboswitch-Mediated Gene Regulation: Novel RNA Architectures Dictate Gene Expression Responses. *Annual Review of Microbiology* 70: 361-374.
8. Baker JL, Sudarsan N, Weinberg Z, Roth A, Stockbridge RB, Breaker RR (2012). Widespread genetic switches and toxicity resistance proteins for fluoride. *Science* 335 (6065): 233-235.
9. Cromie MJ, Shi Y, Latifi T, Groisman EA (2006). An RNA Sensor for Intracellular Mg²⁺. *Cell* 125(1): 71-84.
10. Dann CE, Wakeman CA, Sieling CL, Baker SC, Irnov I, Winkler WC (2007). Structure and Mechanism of a Metal-Sensing Regulatory RNA. *Cell* 130(5): 878-892.
11. Furukawa K, Ramesh A, Zhou Z, Weinberg Z, Vallery T, Winkler WC, Breaker RR (2015). Bacterial Riboswitches Cooperatively Bind Ni²⁺ or Co²⁺ Ions and Control Expression of Heavy Metal Transporters. *Molecular Cell* 57(6): 1088-1098.

12. Shi Y, Zhao G, Kong W (2014). Genetic analysis of riboswitch-mediated transcriptional regulation responding to Mn²⁺ in Salmonella. *Journal of Biological Chemistry* 289(16): 11353–11366.
13. Blount KF, Megyola C, Plummer M, Osterman D, O'Connell T, Aristoff P, Quinn C, Chrusciel RA, Poel TJ, Schostarez HJ, Stewart CA, Walker DP, Wuts PG, Breaker RR (2015). Novel riboswitch-binding flavin analog that protects mice against *Clostridium difficile* infection without inhibiting cecal flora. *Antimicrobial Agents and Chemotherapy* 59(9): 5736-5746.
14. Howe JA, Wang H, Fischmann TO, Balibar CJ, Xiao L, Galgoci AM, Malinverni JC, Mayhood T, Villafania A, Nahvi A, Murgolo N, Barbieri CM, Mann PA, Carr D, Xia E, Zuck P, Riley D, Painter RE, Walker SS, Sherborne B, de Jesus R, Pan W, Plotkin MA, Wu J, Rindgen D, Cummings J, Garlisi CG, Zhang R, Sheth PR, Gill CJ, Tang H, Roemer T (2015). Selective small-molecule inhibition of an RNA structural element. *Nature* 526: 672-677.
15. Howe JA, Xiao L, Fischmann TO, Wang H, Tang H, Villafania A, Zhang R, Barbieri CM, Roemer T (2016). Atomic resolution mechanistic studies of ribocil: A highly selective unnatural ligand mimic of the *E. coli* FMN riboswitch. *RNA Biology* 13(10): 946–954.
16. Lee ER, Blount KF, Breaker RR (2009). Roseoflavin is a natural antibacterial compound that binds to FMN riboswitches and regulates gene expression. *RNA Biology* 6(2): 187-194.
17. Ott E, Stolz J, Lehmann M, Mack M (2009). The RFN riboswitch of *Bacillus subtilis* is a target for the antibiotic roseoflavin produced by *Streptomyces davawensis*. *RNA Biology* 6(3): 276-280.
18. Blount KF, Breaker RR (2006). Riboswitches as antibacterial drug targets. *Nature Biotechnology* 24: 1558-1564.
19. Peselis A, Serganov A (2014). Themes and variations in riboswitch structure and function. *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms* 1839(10): 908-918.

20. Ottink OM, Rampersad SM, Tessari M, Zaman GJR, Heus HA, Wijmenga SS (2007). Ligand-induced folding of the guanine-sensing riboswitch is controlled by a combined predetermined-induced fit mechanism. *RNA* 13(12): 2202-2212.
21. Reining A, Nozinovic S, Schlepckow K, Buhr F, Fürtig B, Schwalbe H (2013). Three-state mechanism couples ligand and temperature sensing in riboswitches. *Nature* 499: 355-359.
22. Hollands K, Proshkin S, Sklyarova S, Epshtein V, Mironov A, Nudler E, Groisman EA (2012). Riboswitch control of Rho-dependent transcription termination. *Proceedings of the National Academy of Sciences of the United States of America* 109(14): 5376-5381.
23. Chen AGY, Sudarsan N, Breaker RR (2011). Mechanism for gene control by a natural allosteric group I ribozyme. *RNA* 17(11): 1967-1972.
24. Bocobza S, Adato A, Mandel T, Shapira M, Nudler E, Aharoni A (2007). Riboswitch-dependent gene regulation and its evolution in the plant kingdom. *Genes and Development* 21(22): 2874–2879.
25. Wachter A, Tunc-Ozdemir M, Grove BC, Green PJ, Shintani DK, Breaker RR (2007). Riboswitch control of gene expression in plants by splicing and alternative 3' end processing of mRNAs. *Plant Cell* 19(11):3437-3450.
26. Antunes D, Jorge NAN, Caffarena ER, Passetti F (2018). Using RNA sequence and structure for the prediction of riboswitch aptamer: A comprehensive review of available software and tools. *Frontiers in Genetics* 8: 231.
27. Kubodera T, Watanabe M, Yoshiuchi K, Yamashita N, Nishimura A, Nakai S, Gomi K, Hanamoto H (2003). Thiamine-regulated gene expression of *Aspergillus oryzae* thiA requires splicing of the intron containing a riboswitch-like domain in the 5'-UTR. *FEBS Letters* 555(3): 1335-1338.
28. Li S, Breaker RR (2013). Eukaryotic TPP riboswitch regulation of alternative splicing involving long-distance base pairing. *Nucleic Acids Research* 41(5): 3022-3031.
29. Croft MT, Moulin M, Webb ME, Smith AG (2007). Thiamine biosynthesis in algae is regulated by riboswitches. *Proceedings of the National Academy of Sciences of the United States of America* 104(52): 20770-20775.

30. George Mason University (2023). NIH Working Definition Of Bioinformatics And Computational Biology [online]. Available from: <http://www.binf.gmu.edu/jafri/math6390-bioinformatics/>. [Accessed 28 February 2023].
31. TÜBA Açık Ders Malzemeleri Portalı (2023). Veri Tabanı Yönetim Sistemleri [online]. Available from: https://acikders.tuba.gov.tr/pluginfile.php/697/mod_resource/content/2/. [Accessed 28 February 2023].
32. Yazağan A. Biyolojik Veritabanları ve Biyoinformatik Analiz Araçları [online]. Available from: <https://docplayer.biz.tr/6629224-Biyolojik-veritabanlari-ve-biyoinformatik-analiz-araclari.html>. [Accessed 28 February 2023].
33. Bioinforange Platformu (2023). Güncel Biyoinformatik Veritabanları ve Yazılım Listesi [online]. Available from: <https://www.bioinforange.com/bioinforeviews/biyoinformatik-veri-tabanlari/biyoinformatik-veri-tabanlari-ve-yazilim-listesi-guncel/>. [Accessed 28 February 2023].
34. Erişime Açık MySQL Veri Tabanı Bölümü (2023). Rfam Yardım Dokümanları [online]. Available from: <https://docs.rfam.org/en/latest/database.html#public-mysql-database>. [Accessed 28 February 2023].
35. Rfam: The RNA families database (2023). Rfam [online]. Available from: <https://rfam.org/>. [Accessed 28 February 2023].
36. Index of /pub/databases/Rfam/CURRENT (2023). Rfam Ftp source [online]. Available from: <https://ftp.ebi.ac.uk/pub/databases/Rfam/CURRENT/>. [Accessed 28 February 2023].
37. Aileler Nasıl Oluşturuldu Bölümü (2023). Rfam Yardım Dokümanları [online]. Available from: <https://docs.rfam.org/en/latest/building-families.html>. [Accessed 28 February 2023].
38. Wray GA (2007). The evolutionary significance of cis-regulatory mutations. *Nature Reviews Genetics* 8: 206-216.

39. Hüttenhofer A, Schattner P, Polacek N (2005). Non-coding RNAs: Hope or hype? Trends in Genetics 21(5): 289-297.
40. Klan Bölümü (2023). Rfam Yardım Dokümanları [online]. Available from: <https://docs.rfam.org/en/latest/glossary.html#clan>. [Accessed 28 February 2023].
41. Aile Bölümü (2023). Rfam Yardım Dokümanları [online]. Available from: <https://docs.rfam.org/en/latest/glossary.html#family>. [Accessed 28 February 2023].
42. Kovaryans Modeli Bölümü (2023). Rfam Yardım Dokümanları [online]. Available from: <https://docs.rfam.org/en/latest/glossary.html#covariance-model-cm>. [Accessed 28 February 2023].
43. Nawrocki EP, Eddy SR (2013). Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics 29(22): 2933-2935.
44. Tohum Hizalama Bölümü (2023). Rfam Yardım Dokümanları [online]. Available from: <https://docs.rfam.org/en/latest/glossary.html#seed-alignment>. [Accessed 28 February 2023].
45. Tam Hizalama Bölümü (2023). Rfam Yardım Dokümanları [online]. Available from: <https://docs.rfam.org/en/latest/glossary.html#full-alignment>. [Accessed 28 February 2023].
46. Nawrocki EP, Burge SW, Bateman A, Daub J, Eberhardt RY, Eddy SR, Floden EW, Gardner PP, Jones TA, Tate J, Finn RD (2015). Rfam 12.0: Updates to the RNA families database. Nucleic Acids Research 43(D): 130-137.
47. Lee JC, Gutell RR (2004). Diversity of base-pair conformations and their occurrence in rRNA structure and RNA structural motifs. Journal of Molecular Biology 344(5): 1225-1249.
48. Baloğlu MC, Bayraç AT, Sadi G, Bayraç C (2012). Moleküler Biyoloji. Genişletilmiş İkinci Baskı. Nobel Akademik Yayıncılık, Ankara; Sayfa: 309-340.
49. GFF/GTF File Format (2023). GFF/GTF File Format-Definition and supported options [online]. Available from: <https://www.ensembl.org/info/website/upload/gff.html>. [Accessed 28 February 2023].
50. wwPDB consortium. (2019). Protein Data Bank: The single global archive for 3D macromolecular structure data. Nucleic Acids Research 47(D1): 520-528.

51. IDX file format (2023). Phonetic Sciences, Amsterdam [online]. Available from: https://www.fon.hum.uva.nl/praat/manual/IDX_file_format.html. [Accessed 28 February 2023].
52. Lodish HF (2011). Molecular Cell Biology. Moleküler Hücre Biyolojisi. 6th ed. Çeviri Editörleri: Geçkıl H, Özmen M, Yeşilada Ö, Palme Yayınevi, Ankara, 123.
53. Carroll H, Ridge P, Clement M, Snell Q (2006). Effects of gap open and gap extension penalties [online]. Available from: <https://scholarsarchive.byu.edu/cgi/viewcontent.cgi?article=1289&context=facpub>. [Accessed 28 February 2023].
54. Johnson M, Zaretskaya I, Raytselis Y, Merezuk Y, McGinnis S, Madden TL (2008). NCBI BLAST: a better web interface. Nucleic acids research 36(W): 5-9.
55. Galtier N, Lobry JR (1997). Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes. Journal of Molecular Evolution 44(6): 632-636.
56. Abril JF, Castellano S (2019). Genome Annotation. Ranganathan S, Gribskov M, Nakai K, Schönbach C (Ed.), Encyclopedia of Bioinformatics and Computational Biology, 195-209. <https://www.sciencedirect.com/science/article/pii/B9780128096338202264>
57. Madeira F, Pearce M, Tivey ARN, Basutkar P, Lee J, Edbali O, Madhusoodanan N, Kolesnikov A, Lopez R (2022). Search and sequence analysis tools services from EMBL-EBI in 2022. Nucleic Acids Research 50(W1): 276-279.
58. Protein structure prediction (2023). Wikipedia The Free Encyclopedia [online]. Available from: https://en.wikipedia.org/wiki/Protein_structure_prediction. [Accessed 28 February 2023].
59. Sunburst Chart (2023). AnyChart [online]. Available from: <https://www.anychart.com/chartopedia/chart-type/sunburst-chart/>. [Accessed: 28 February 2023].
60. Van Rossum G, Drake FL, Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, Wieser E, Taylor J, Berg S, Smith N, Kern R, Picus M, Hoyer S, van Kerkwijk M, Brett M, Haldane A, Del Río J, Wiebe M, Peterson P,

- Gérard-Marchant P, Sheppard K, Reddy T, Weckesser W, Abbasi H, Gohlke C, Oliphant T (2009). Python 3 Reference Manual. Nature 585: 357-362.
61. JetBrains/intellij-community (2023). PyCharm Community Edition, Github source [online]. Available from: <https://github.com/JetBrains/intellij-community/tree/master/python>. [Accessed 28 February 2023].
 62. Anaconda Software Distribution (2023). Anaconda Documentation [online]. Available from: <https://docs.anaconda.com/>. [Accessed 28 February 2023].
 63. Python Package Index (2023). PyPI Documentation [online]. Available from: <https://pypi.org/project/documentation/>. [Accessed 28 February 2023].
 64. Uyguçgil H (2018). Veritabanı Kuramı. Uyguçgil H (Ed.), Konumsal Veritabanı-I, 30-34. <https://ets.anadolu.edu.tr/storage/nfs/CBS113U/ebook/CBS113U-17V1S1-8-0-1-SV1-ebook.pdf>
 65. Benslimane D, Dustdar S, Sheth A (2008). Services mashups: The new generation of web applications. IEEE Internet Computing 12(5): 13-15.
 66. API (2023). Wikipedia The Free Encyclopedia [online]. Available from: <https://en.wikipedia.org/wiki/api>. [Accessed 8 April 2023].
 67. What is a Data-pipeline? (2023). Data Pipelines [online]. Available from: <https://www.datapipelines.com/blog/what-is-a-data-pipeline/>. [Accessed: 28 February 2023].
 68. Regular expression (2023). Wikipedia The Free Encyclopedia [online]. Available from: https://en.wikipedia.org/wiki/Regular_expression. [Accessed 8 April 2023].
 69. TRUBA (2023). Türk Ulusal Bilim e-Altyapısı [online]. Available from: <https://www.truba.gov.tr/>. [Accessed: 28 February 2023].
 70. Slurm Workload Manager (2023). SchedMD [online]. Available from: <https://slurm.schedmd.com/quickstart.html>. [Accessed: 28 February 2023].
 71. Ligand (2023). Wikipedia The Free Encyclopedia [online]. Available from: <https://en.wikipedia.org/wiki/ligand>. [Accessed 8 April 2023].
 72. Rang HP (2006). The receptor concept: Pharmacology's big idea. British Journal of Pharmacology 147(S1): 9-16.

73. Kitchen DB, Decornez H, Furr JR, Bajorath J (2004). Docking and scoring in virtual screening for drug discovery: Methods and applications. *Nature Reviews Drug Discovery* 3: 935-949.
74. Hassan NM, Alhossary AA, Mu Y, Kwoh CK (2017). Protein-Ligand Blind Docking Using QuickVina-W with Inter-Process Spatio-Temporal Integration. *Scientific Reports* 7(1): 15451.
75. Morris GM, Ruth H, Lindstrom W, Sanner MF, Belew RK, Goodsell DS, Olson AJ (2009). Software news and updates AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *Journal of Computational Chemistry* 30(16): 2785-2791.
76. Trott O, Olson AJ (2010). AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading. *Journal of Computational Chemistry* 31(2): 455-461.
77. Eberhardt J, Santos-Martins D, Tillack AF, Forli S (2021). AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings. *Journal of Chemical Information and Modeling* 61(8): 3891-3898.
78. Nocedal J, Wright SJ (2006). Quasi Newton Methods. Mikosch TV, Resnick SI, Robinson SM (Ed.), *Numerical Optimization*, 136-142. <http://egrcc.github.io/docs/math/numerical-optimization.pdf>
79. Schrödinger L, DeLano W (2020). PyMOL [online]. Available from: <http://www.pymol.org/pymol>. [Accessed: 6 April 2023].
80. AutoDock 4.2 User Guide (2023). Scripps Research [online]. Available from: https://autodock.scripps.edu/wp-content/uploads/sites/56/2021/10/AutoDock426_UserGuide.pdf. [Accessed: 28 February 2023].
81. Mol2 File Format Explained for Beginners (2023). ChemicBook, Chemistry Meets Data Science [online]. Available from: <https://chemicbook.com/2021/02/20/mol2-file-format-explained-for-beginners-part-2.html>. [Accessed: 28 February 2023].

82. Rester U (2008). From virtuality to reality-Virtual screening in lead discovery and lead optimization: A medicinal chemistry perspective. *Current Opinion in Drug Discovery and Development* 11(4): 559-568.
83. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ (2012). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews* 64: 4-17.
84. Lipinski CA (2004). Lead- and drug-like compounds: The rule-of-five revolution. *Drug Discovery Today: Technologies* 1(4): 337-341.
85. Maggiora G, Vogt M, Stumpfe D, Bajorath J (2014). Molecular similarity in medicinal chemistry. *Journal of Medicinal Chemistry* 57(8): 3186-3204.
86. Gerald B (2018). A Brief Review of Independent, Dependent and One Sample t-test. *International Journal of Applied Mathematics and Theoretical Physics* 4(2): 50-54.
87. Khurana D, Koli A, Khatter K, Singh S (2023). Natural language processing: state of the art, current trends and challenges. *Multimedia Tools and Applications* 82: 3713-3744.
88. Martinez AM, Kak AC (2001). PCA versus LDA. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23(2): 228-233.
89. Halvey MJ, Keane MT (2007). An Assessment of Tag Presentation Techniques. In: *Proceedings of the 16th International Conference on World Wide Web - WWW '07*, New York USA, 8-12 May 2007, 1313.
90. Hearst M (2003). What Is Text Mining?. SIMS, UC Berkeley [online]. Available from: <https://people.ischool.berkeley.edu/~hearst/text-mining.html>. [Accessed: 28 February 2023].
91. Bartko JJ (1966). The intraclass correlation coefficient as a measure of reliability. *Psychological reports* 19(1): 3-11.
92. Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, de Hoon MJ (2009). Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25(11): 1422-1423.

93. Madeira F, Park YM, Lee J, Buso N, Gur T, Madhusoodanan N, Basutkar P, Tivey ARN, Potter SC, Finn RD, Lopez R (2019). The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Research* 47(W1): 636-641.
94. Rangwala SH, Kuznetsov A, Ananiev V, Asztalos A, Borodin E, Evgeniev V, Joukov V, Lotov V, Pannu R, Rudnev D, Shkeda A, Weitz EM, Schneider VA (2021). Accessing NCBI data using the NCBI sequence viewer and genome data viewer (GDV). *Genome Research* 31(1): 159-169.
95. McKinney W (2010). Data Structures for Statistical Computing in Python. Proceedings of the 9th Python in Science Conference, Austin Texas, June 28-July 3, 56-61.
96. Sayers E (2012). E-utilities Quick Start Entrez Programming Utilities Help Entrez Programming Utilities Help. *The Journal of Systems and Software* 85(D1): 13-25.
97. PMID vs PMCID: What's the Difference?. NIH Staff [online]. Available from: <https://nexus.od.nih.gov/all/2015/08/31/pmid-vs-pmcid-whats-the-difference> [Accessed: 7 March 2023].
98. Comeau DC, Wei CH, Islamaj Doğan R, Lu Z (2019). PMC text mining subset in BioC: About three million full-text articles and growing. *Bioinformatics* 35(18): 3533-3535.
99. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay É (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12: 2825-2830.
100. Anlamli Sekans Dosyalarının Temin Edilmesi (2023). Rfam Yardım Dokümanları [online]. Available from: <https://docs.rfam.org/en/latest/database.html#retrieve-all-rat-sequence-coordinates-annotated-with-rfam-families>. [Accessed: 28 February 2023].
101. PDB-FTP Archive over HTTP (2023). PDB-FTP Archive over HTTP [online]. Available from: https://ftp.rcsb.org/pub/pdb/derived_data/index/. [Accessed: 28 February 2023].
102. ZINC15 Database (2023). Category:ZINC15 [online]. Available from: <https://wiki.docking.org/index.php/Category:ZINC15>. [Accessed: 7 April 2023].
103. Resources-DISI (2023). Wiki docking electronic resources [online]. Available from: <https://wiki.docking.org/index.php/Resources>. [Accessed: 7 April 2023].

104. Irwin JJ, Gaskins G, Sterling T, Mysinger MM, Keiser MJ (2018). Predicted Biological Activity of Purchasable Chemical Space. *Journal of Chemical Information and Modeling* 58(1): 148-164.
105. The Open Babel Package, version 2.3.1 (2023). Open Babel: The Open Source Chemistry Toolbox [online]. Available from: <http://openbabel.org>. [Accessed: 28 February 2023].
106. O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR (2011). Open Babel: An Open chemical toolbox. *Journal of Cheminformatics* 3: 33.
107. Garcia-Ortegón M, Simm GNC, Tripp AJ, Hernández-Lobato JM, Bender A, Bacallado S (2022). DOCKSTRING: Easy molecular docking yields better benchmarks for ligand design. *J Chem Inf Model* 62: 3486-3502.
108. Serganov A, Yuan YR, Pikovskaya O, Polonskaia A, Malinina L, Phan AT, Hobartner C, Micura R, Breaker RR, Patel DJ (2004). Structural basis for discriminative regulation of gene expression by adenine- and guanine-sensing mRNAs. *Chemistry and Biology* 11(12): 1729-1741.
109. Introduction to in silico docking (2023). Ross G, SBCB, University of Oxford 2012 [online]. Available from: <https://sbc.bioch.ox.ac.uk/users/greg/teaching/docking-2012.html>. [Accessed: 7 March 2023].
110. Sanner MF (1999). Python: A programming language for software integration and development. *Journal of Molecular Graphics and Modelling* 17(1): 57-61.
111. Liberman JA, Suddala KC, Aytenfisu A, Chan D, Belashov IA, Salim M, Mathews DH, Spitale RC, Walter NG, Wedekind JE (2015). Structural analysis of a class III preQ1 riboswitch reveals an aptamer distant from a ribosome-binding site regulated by fast dynamics. *Proceedings of the National Academy of Sciences of the United States of America* 112(27): 3485-3494.
112. Butler EB, Xiong Y, Wang J, Strobel SA (2011). Structural basis of cooperative ligand binding by the glycine riboswitch. *Chemistry and Biology* 18(3): 293-298.
113. Serganov A, Yuan YR, Pikovskaya O, Polonskaia A, Malinina L, Phan AT, Hobartner C, Micura R, Breaker RR, Patel DJ (2004). Structural basis for

- discriminative regulation of gene expression by adenine- and guanine-sensing mRNAs. *Chemistry and Biology* 11(12): 1729-1741.
114. Dunstan MS (2019). X-ray crystal structure of the tetrahydrofolate riboswitch aptamer bound to 5-deazatetrahydropterin [online]. Available from: https://www.rcsb.org/pdb?id=pdb_00006q57. [Accessed: 28 February 2023].
 115. Weickhmann AK, Keller H, Wurm JP, Streibitzer E, Juen MA, Kremser J, Weinberg Z, Kreutz C, Duchardt-Ferner E, Wöhnert J (2019). The structure of the SAM/SAH-binding riboswitch. *Nucleic Acids Research* 47(5): 2654-2665.
 116. RDKit: Open-source cheminformatics (2023). RDKit [online]. Available from: <https://www.rdkit.org>. [Accessed: 7 March 2023].
 117. Kochnev Y, Hellemann E, Cassidy KC, Durrant JD (2020). Webina: An open-source library and web app that runs AutoDock Vina entirely in the web browser. *Bioinformatics* 36(16): 4513-4515.
 118. Shrout PE, Fleiss JL (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin* 86(2): 420-428.
 119. Bujang MA, Baharum N (2017). A simplified guide to determination of sample size requirements for estimating the value of intraclass correlation coefficient: A review. *Archives of Orofacial Sciences* 12(1): 1-11.
 120. Cicchetti D V. (1994). Guidelines, Criteria, and Rules of Thumb for Evaluating Normed and Standardized Assessment Instruments in Psychology. *Psychological Assessment* 6(4): 284-290.
 121. Frequently Asked Questions (2023). AutoDock Vina [online]. Available from: <https://autodock-vina.readthedocs.io/en/latest/faq.html>. [Accessed: 7 March 2023].
 122. Bross MH, Soch K, Morales R, Mitchell RB (2007). *Vibrio vulnificus* infection: Diagnosis and treatment. *American Family Physician* 76(4): 539-544.
 123. Compound Libraries for High Throughput/Content Screening (2023). FDA-approved Drug Library [online]. Available from: <https://www.selleckchem.com/screening/fda-approved-drug-library.html>. [Accessed: 7 March 2023].

124. Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, Li Q, Shoemaker BA, Thiessen PA, Yu B, Zaslavsky L, Zhang J, Bolton EE (2023). PubChem 2023 update. *Nucleic Acids Research* 51(D): 1373-1380.
125. Cornwell DG (1962). *Chemistry of the amino acids*, 3 volumes (Greenstein, Jesse P.; Winitz, Milton). *Journal of Chemical Education* 39(4): 217-218.
126. Chow SL, Rohan C, Morris AAM (2003). Case Report: Rhabdomyolysis in Glutaric Aciduria Type I. *Journal of Inherited Metabolic Disease* 26: 711-712.
127. Strauss KA, Puffenberger EG, Robinson DL, Morton DH (2003). Type I glutaric aciduria, part 1: Natural history of 77 patients. *American Journal of Medical Genetics - Seminars in Medical Genetics* 121C(1): 38-52.
128. Kölker S, Koeller DM, Okun JG, Hoffmann GF (2004). Pathomechanisms of Neurodegeneration in Glutaryl-CoA Dehydrogenase Deficiency. *Annals of Neurology* 55(1): 7-12.
129. Boy N, Mühlhausen C, Maier EM, Heringer J, Assmann B, Burgard P, Dixon M, Fleissner S, Greenberg CR, Harting I, Hoffmann GF, Karall D, Koeller DM, Krawinkel MB, Okun JG, Opladen T, Posset R, Sahm K, Zschocke J, Kölker S (2017). Proposed recommendations for diagnosing and managing individuals with glutaric aciduria type I: second revision. *Journal of Inherited Metabolic Disease* 40(1): 75-101.
130. Boy N, Garbade SF, Heringer J, Seitz A, Kölker S, Harting I (2019). Patterns, evolution, and severity of striatal injury in insidious- vs acute-onset glutaric aciduria type 1. *Journal of Inherited Metabolic Disease* 42(1): 117-127.
131. Strauss KA, Morton DH (2003). Type I glutaric aciduria, part 2: A model of acute striatal necrosis. *American Journal of Medical Genetics - Seminars in Medical Genetics* 121(C1): 53-70.
132. Pradhan P, Soni NK, Chaudhary L, Mujwar S, Pardasani KR (2015). In-silico prediction of riboswitches and design of their potent inhibitors for H1N1, H2N2 and H3N2 strains of influenza virus. *Biosciences Biotechnology Research Asia* 12(3): 2173-2186.
133. Pandey M, Pardasani KR (2016). In Silico Prediction of Riboswitch and Design of its Inhibitor for Salmonella enterica. *Drug Designing: Open Access* [online]. Available

from: <https://www.longdom.org/open-access/in-silico-prediction-of-riboswitch-and-design-of-its-inhibitor-for-salmonella-enterica-2169-0138-1000136.pdf>. [Accessed: 7 March 2023].

134. Mujwar S, Pardasani KR (2015). Prediction of riboswitch as a potential drug target for infectious diseases: An insilico case study of anthrax. *Journal of Medical Imaging and Health Informatics* 5(1): 7-16.





EKLER

EK 1. Tez Çalışması Kapsamında Elde Edilen Dijital Veriler

Bu çalışma kapsamında elde veriler tez metni içerisinde yer alan ek kısmında verilemeyecek kadar büyük boyutludur. Bu nedenle çalışma sürecinde elde edilen tüm verilere ve bu verilerden elde edilen rapor ve bulgu içeriklerine <https://sites.google.com/view/nbzphdthesis/ana-sayfa> adresinden ulaşılabilir.



- 2 Zihni NB, Turhan K, Çan MA (2013). Tıp Eğitime Yönelik Olarak Hazırlanan Uzaktan Eğitim Sitesindeki İnsan Bilgisayar Etkileşiminin İncelenmesi. Sigma Mühendislik ve Fen Bilimleri Dergisi 5(1):15-24.

BİLDİRİLER

- 1 Zihni NB, Turhan K, Aydın Kasap Z (2018). Karnofsky Performans Ölçeğinin Çeşitli Kanseri Kohortlarının Sağkalımı Üzerine Etkisinin İncelenmesi. XX. Ulusal ve III. Uluslararası Biyoistatistik Kongresi, Trabzon, 26-29 Ekim 2018, 10-12.
- 2 Zihni NB, Uzun Ö, Yeginoğlu G, Kalkışım ŞN, Ertemoğlu Öksüz C (2017). Çoklu Uyum Analizi ve Bir Uygulaması. XIX. Ulusal ve II. Uluslararası Biyoistatistik Kongresi, Antalya, 25-28 Ekim 2017, 61-62.
- 3 Zihni NB, Turhan K, Çan MA (2012). Tıp Eğitiminde Web Tabanlı Uzaktan Eğitim Uygulaması ve Öğrenci Başarısına Etkisinin İncelenmesi. IX. Ulusal Tıp Bilişimi Kongresi, Antalya, 15-18 Kasım 2012, 128-135.