**RNA 3D Yapı Tahmini Yarışması - Kapsamlı Özet**

**Giriş**

RNA 3D yapı tahmini, RNA moleküllerinin fonksiyonlarını anlamak ve ilaç tasarımı gibi uygulamalar açısından kritik öneme sahiptir. Bu projede, RNA dizilerinden üç boyutlu yapılarını tahmin etmek için çeşitli stratejiler geliştirilmiş ve test edilmiştir. En yüksek skoru 0.205 ile "İnce Ayarlı RNA Uzunluğuna Adaptif Parametre" stratejisinde elde ettik. Tüm çalışmalarda RandomForest modelleri temel alınmış ve TM-score metriği ile stratejilerin başarısı değerlendirilmiştir. Nature Methods makalesinde vurgulanan dört temel zorluğa (Watson-Crick olmayan baz çiftlerinin doğru tahmini, sarmallar arasındaki koaksiyel istiflemenin doğru konumlandırılması, dolaşıklıkların önlenmesi ve helikal eksenler arasındaki açıların doğru modellenmesi) odaklanan yaklaşımlar denenmiştir.

**Denenen Stratejiler (Düşük Skordan Yüksek Skora)**

**1. CNN-Entegrasyonlu Hibrit Model (0.089)**

Bu strateji, RandomForest yaklaşımını basit bir CNN modeli ile desteklemeyi hedefledi. Ancak beklenmedik şekilde en düşük skoru verdi:

* **Metodoloji**: RandomForest tahminleri üzerine CNN modelinden gelen tahminlerin entegrasyonu
* **Başarısızlık Nedeni**: Nature Methods makalesinde de belirtildiği gibi, RNA yapı tahmininde derin öğrenme yaklaşımları henüz geleneksel yaklaşımlar kadar başarılı değil
* **Temel Bulgu**: RNA yapısının karmaşıklığı ve veri kısıtlılığı, derin öğrenmenin protein yapı tahminindeki kadar başarılı olmasını engelliyor

**2. XGBoost Stratejisi (0.129)**

RandomForest yerine daha güçlü XGBoost regresyon modellerini kullanmayı hedefleyen strateji:

* **Metodoloji**: 300 ağaç, 0.05 öğrenme oranı, 6 maksimum derinlik ile XGBoost regresyon modelleri
* **Eklenen Özellikler**: Dinükleotid kompozisyonu, RNA dizilerindeki motif ve paternleri yakalamak için
* **Sonuç**: Beklenenin aksine daha düşük performans, aşırı uyum (overfitting) problemi gözlendi

**3. Ensemble Motif-Koaksiyel (0.142)**

Karmaşık bir yaklaşım olarak tasarlanan bu strateji beklenenden çok düşük performans gösterdi:

* **Stratejinin Bileşenleri**: Motif tanımlama, koaksiyel istifleme tespiti, yapılar arası karmaşık korelasyonlar
* **Başarısızlık Nedenleri**:
  + Aşırı karmaşıklık (çok sayıda parametre ve varyasyon mekanizması)
  + Motif tanımlama sorunları
  + Yapılar arası karmaşık korelasyonların hata birikimini artırması

**4. Gelişmiş RandomForest Parametreleri (0.149)**

RandomForest hiperparametrelerinin optimize edilmesi:

* **Değişiklikler**:
  + Ağaç sayısı artırıldı (100 → 200)
  + Derinlik ve bölünme kriterlerinde düzenlemeler yapıldı
* **Sonuç**: Performans düşüşü - model eğitim verisine aşırı uyum gösterdi

**5. RMSD Odaklı Motif Geliştirilmiş (0.174) ve Ultra Agresif Extreme Varyasyon (0.175)**

Bu stratejilerde farklı yaklaşımlar denendi:

* **RMSD Odaklı Yaklaşım**: Kabsch algoritması ile RMSD optimizasyonu
* **Ultra Agresif Varyasyon**: Çok yüksek gürültü seviyeleri (0.35, 0.60, 0.95, 0.85) × 10 ve farklı dağılımlar
* **Sonuç**: Daha agresif yaklaşımlar daha düşük performansa yol açtı

**6. Fizik Kısıtlamalı Korelasyonlu Gürültü (0.178)**

Fizik kurallarının RNA modeline entegrasyonunu hedefleyen strateji:

* **Kısıtlamalar**:
  + Atom çakışmalarını önleme (minimum 3.5 Å mesafe)
  + Bağ uzunluğu koruması (5.5-6.5 Å)
* **Sonuç**: Aşırı kısıtlamalar varyasyon çeşitliliğini azalttı ve doğal RNA esnekliğini sınırladı

**7. Temel RandomForest + Basit Rastgele Varyasyonlar (0.179)**

İlk baseline denemesi beklenenden daha iyi sonuç verdi:

* **Metodoloji**: 100 ağaçlı basit RandomForest modeli ve artan gürültü seviyelerinde rastgele varyasyonlar
* **Sonuç**: Basitliğine rağmen oldukça başarılı bir yaklaşım olduğu kanıtlandı

**8. Biyolojik Varyasyonlar ve Fizik Kısıtlamaları (0.181)**

RNA'nın doğal yapısını taklit eden varyasyonlar:

* **Biyolojik Faktörler**: RNA uçlarında esneklik, rotasyon, bükülme
* **Fiziksel Faktörler**: Atom mesafeleri ve çakışmaları için kısıtlamalar
* **Sonuç**: Tohum optimizasyonu kadar etkili değil, ama makul sonuçlar verdi

**9. Hibrit Optimizasyon ve Seed Optimizasyonu (0.183)**

Farklı yaklaşımların bir araya getirildiği stratejiler:

* **Hibrit Optimizasyon**: SVD Procrustes Hizalama + Mesafe Matrisi + t-dağılımı + Motif-Bazlı Yaklaşım
* **Seed Optimizasyonu**: Farklı rastgele tohum değerlerinin test edilmesi
* **Temel Bulgu**: Seed 123 değeri tutarlı şekilde en iyi sonuçları verdi

**10. Student's t-Dağılımı ve Geliştirilmiş Seed 123 Extreme Varyasyon (0.190)**

Bu stratejilerde farklı dağılımlar ve varyasyon kombinasyonları denendi:

* **Student's t-Dağılımı**: Her yapı için farklı serbestlik derecesi (df=5.0→4.0→3.0→2.0)
* **Geliştirilmiş Seed 123**: Geniş gürültü spektrumu ve nükleotid tipine göre adaptif faktörler
* **Sonuç**: Ağır kuyruklu dağılımlar nadir konformasyonları daha iyi yakaladı

**11. Çeşitli Ensemble Yaklaşımları (0.191)**

Farklı ensemble yaklaşımları benzer skorlar verdi:

* **Ensemble Yaklaşımı**: İstatistiksel dağılım çeşitlendirmesi ve adaptif faktörler
* **Tohum ve Dağılım Optimizasyonu**: Her yapı için farklı bir tohum değeri ve çeşitli dağılımlar
* **Motif-Bazlı Yapılar Arası Korelasyonlu Gürültü**: En başarılı üç stratejinin birleşimi

**12. RNA Motif-Bazlı Varyasyon Stratejisi (0.192)**

RNA'nın yapısal motiflerini tahmin eden algoritma kullanıldı:

* **Motif Tipleri**: Stem, loop, bulge, hairpin
* **Esneklik Faktörleri**: Stem için düşük (0.55-0.7), loop/bulge için yüksek (1.3-2.0)
* **Koordinasyon**: Aynı motifteki nükleotidlerin benzer hareket etmesi sağlandı

**13. Yapılar Arası Korelasyon Stratejisi (0.193)**

Her yapının bir önceki yapıdan etkilendiği bir model tasarlandı:

* **Etkileşim**: Yapı 2→3→4→5 şeklinde etkilenim
* **Korelasyon Ağırlıkları**: 0.3, 0.35, 0.4
* **Sonuç**: RNA konformasyon geçişlerindeki devamlılık iyi modellenmiş

**14. Orijinal Korelasyonlu Gürültü (0.194)**

Nükleotidlerin koordineli hareket etmesini sağlayan temel yaklaşım:

* **Metodoloji**:
  + Her nükleotid için X, Y, Z eksenleri arasında korelasyonlu hareket
  + Seed değeri: 123
  + Farklı dağılımlar (Laplace, Normal) ve gürültü seviyeleri
* **Korelasyon Ağırlıkları**: 0.7, 0.7, 0.75, 0.8
* **Sonuç**: Uzun süre en yüksek skoru veren strateji oldu

**15. Motif-Temelli Adaptif Gürültü (0.195)**

Her nükleotid için RNA yapısal motiflerini tanımlayıp, motif tiplerine göre farklı esneklik faktörleri uygulandı:

* **Faktörler**: Stem: 0.7, Bulge: 1.5, Hairpin: 1.3, Loop: 1.2, Terminal: 1.8
* **Metodoloji**: Baz çiftlerini ve sekonder yapıyı tespit eden bir algoritma
* **Sonuç**: Daha gerçekçi RNA konformasyonları oluşturuldu

**16. Kabsch-Geliştirilmiş Korelasyonlu Gürültü (0.196)**

3D rotasyon matrisleri ve matematiksel transformasyonlar kullanıldı:

* **Transformasyonlar**: Dönme, uzama, makaslama, Kabsch hizalama
* **Dağılımlar**: Farklı istatistiksel dağılımlar ve gürültü seviyeleri
* **Metematiksel Yaklaşım**: RNA yapılarını en iyi şekilde hizalamak için Kabsch algoritması

**17. Temel Termal Ensemble Kabsch (0.198)**

Moleküler termodinamik prensiplerini taklit ederek farklı "sıcaklık" seviyelerinde RNA konformasyonları oluşturuldu:

* **Yapı Oluşturma**:
  + Yapı 1: Temel RandomForest tahminleri
  + Yapı 2 (Düşük Sıcaklık - 0.25×10): Az hareket, Laplace dağılımı
  + Yapı 3 (Orta Sıcaklık - 0.45×10): Orta seviye, karışık dağılım
  + Yapı 4 (Yüksek Sıcaklık - 0.75×10): Büyük hareketler, t-dağılımı
  + Yapı 5 (Hibrit - 0.65×10): Önceki yapıların ağırlıklı ortalaması
* **Adaptif Faktörler**: Nükleotid tipine ve RNA motiflerine göre uyarlanmış faktörler

**18. Biyolojik Filtreleme (0.199)**

RNA moleküllerinin fiziksel özelliklerini modele dahil eden en yüksek skorlu stratejilerden biri:

* **Kısıtlamalar**:
  + Atom çakışmalarının önlenmesi (min. 3.8 Å)
  + Bağ uzunluğu koruması (∼6.0 Å)
  + Motif bazlı düzeltmeler
* **İteratif Düzeltme**: 3 iterasyonda azalan şiddette kısıtlamalar (0.8→0.5→0.3)

**19. RNA Uzunluğuna Adaptif Parametre Yaklaşımı (0.199 → 0.205)**

En yüksek skoru veren bu strateji, RNA uzunluğuna göre farklı parametreler kullanmayı hedefledi:

* **Kategoriler**:
  + Kısa RNA'lar (<40 nükleotid): Daha sıkı parameterler
  + Orta RNA'lar (40-120 nükleotid): Standart parameterler
  + Uzun RNA'lar (>120 nükleotid): Daha esnek parameterler
* **İnce Ayarlar**: Kategori sınırları ve parametre değerleri optimize edildi
* **Sonuç**: Bu strateji en yüksek ve en tutarlı sonuçları verdi (0.205)

**Sonraki Stratejiler ve Sonuçları**

Daha sonra denenen stratejilerde, başarılı olan RNA Uzunluğuna Adaptif Parametre yaklaşımı üzerine ek iyileştirmeler yapılmaya çalışıldı:

**Watson-Crick Olmayan Baz Çiftleri Tespiti (0.205)**

RNA Uzunluğuna Adaptif temeline eklenen bu strateji skoru korusa da iyileştirmedi:

* **Yaklaşım**: Sofistike bir non-WC baz çifti tespit sistemi ve geometrik kısıtlamalar
* **Sonuç**: Ek karmaşıklık eklerken skor artışı sağlamadı

**Koaksiyel İstifleme Optimizasyonu (0.203)**

Heliksler arasındaki koaksiyel istiflemeyi optimize eden bu yaklaşım skorda düşüşe yol açtı:

* **Değişiklikler**: Açı (25°→20°) ve mesafe (7.5Å→6.5Å) eşiklerinin hassaslaştırılması
* **Skordaki Düşüş**: 0.205→0.203 (%0.98 azalma)

**Fiziksel Dolaşıklık Önleme (0.202)**

Dolaşıklıkları önlemek için kompleks bir algoritma eklendi:

* **Metodoloji**: KD-ağacı kullanarak dolaşıklıkları tespit etme ve fiziksel enerji minimizasyonu
* **Sonuç**: Çok fazla dolaşıklık tespit edildi (toplam 53,431), skorda düşüş yaşandı

**Torsiyonel Açı Optimizasyonu (0.200)**

RNA'daki bilinen motifler için torsiyonel açıları optimize etmeyi hedefledi:

* **Metodoloji**: Her motif tipine özgü ideal açı değerleri ve kademeli düzeltme
* **Sonuç**: Strateji karmaşıklığı artırırken skoru düşürdü

**Öğrenilen Dersler ve İçgörüler**

**Stratejik İçgörüler**

1. **Basitlik Güçtür**: Daha karmaşık modeller her zaman daha iyi performans göstermez. En yüksek skorlar genellikle temel modeller ve iyi düzenlenmiş parametre setleriyle elde edildi.
2. **Rastgeleliğin Gücü**: RNA yapısındaki doğal varyasyonları yakalamada rastgele yaklaşımlar etkilidir. Tohum (seed) değeri 123 tutarlı şekilde en iyi sonuçları verdi.
3. **RNA Uzunluğunun Önemi**: Farklı uzunluktaki RNA'lar farklı yapısal davranışlar gösterir ve farklı parametre setleri gerektirir. Test setindeki 12 RNA'nın çoğunluğu (7) uzun RNA'lardı, bu nedenle uzun RNA'lar için yapılan optimizasyonlar skorları daha çok etkiledi.
4. **Biyolojik Filtrelemenin İkili Doğası**: Biyolojik filtreleme yapıyı gerçekçi kılarken, modelin doğal esnekliğini sınırlayabilir. Doğru denge kritik öneme sahiptir.
5. **Korelasyonlu Hareket Modelleme**: RNA moleküllerinin gerçek dünyada korelasyonlu hareket ettiği gerçeği modele yansıtılmalıdır. X, Y, Z koordinatları arasında korelasyon sağlayan yaklaşımlar başarılı olmuştur.

**Teknik Dersler**

1. **Dağılım Seçiminin Önemi**: Farklı istatistiksel dağılımlar (Normal, Laplace, Student's t) farklı senaryolarda avantaj sağlar. Yapı 2'de Laplace dağılımı, diğer yapılarda Gaussian veya Student's t dağılımı daha iyi çalıştı.
2. **RNA Motif Analizi**: RNA yapısal motifleri (stem, loop, bulge) yapı tahmininde önemli faktörlerdir. Stem bölgeleri daha katı, loop ve bulge bölgeleri daha esnek modellenmelidir.
3. **Matematik Temelli Yaklaşımlar**: Kabsch algoritması gibi matematiksel temelli yaklaşımlar güçlü sonuçlar verebilir. Bu tür yaklaşımlar, RNA yapılarının global hizalamasını iyileştirebilir.
4. **Kademeli İlerleme Önemi**: Bir seferde çok fazla değişiklik yapmak yerine, tek tek değişiklikler yapıp etkilerini ölçmek daha etkilidir. Her değişikliğin etkisini ayrı ayrı değerlendirmek kritik öneme sahiptir.
5. **Derin Öğrenme Sınırlamaları**: Nature Methods makalesinde de belirtildiği gibi, derin öğrenme yöntemleri henüz RNA yapı tahmininde geleneksel yaklaşımlar kadar başarılı değildir. CNN entegrasyonu en düşük skoru verdi (0.089).
6. **Optimal Çeşitlilik**: Çeşitliliği artırmak her zaman daha iyi sonuçlar vermeyebilir. Belirli bir noktadan sonra fazla çeşitlilik, faydalı sinyalleri gizleyebilir. Doğru dengeyi bulmak önemlidir.

**Sonuç**

RNA 3D yapı tahmini projesi, birçok farklı stratejinin sistematik olarak değerlendirilmesini sağlamıştır. En yüksek skoru (0.205) elde eden "İnce Ayarlı RNA Uzunluğuna Adaptif Parametre" yaklaşımı, RNA moleküllerinin uzunluğuna göre optimize edilmiş parametreler kullanmanın ve basit ancak iyi ayarlanmış bir modelin etkili olabileceğini göstermiştir. Daha karmaşık stratejiler her zaman daha iyi performans vermemiş, hatta bazı durumlarda skorların düşmesine yol açmıştır.

Nature Methods makalesinde vurgulanan dört temel zorluğu (Watson-Crick olmayan baz çiftleri, koaksiyel istifleme, dolaşıklıklar ve helikal eksen açıları) adresleyen stratejiler denenmiş, ancak bu karmaşık iyileştirmeler beklenenden daha az etki yaratmıştır. RNA yapı tahmini alanında, karmaşık hesaplamalardan önce, biyolojik ve fiziksel gerçeklik prensiplerine bağlı kalmak, daha doğru ve güvenilir sonuçlar elde etmenin anahtarı olarak ortaya çıkmıştır.

Test veri setindeki RNA uzunluk dağılımının (çoğunlukla uzun RNA'lar) sonuçları etkilediği görülmüş ve uzunluğa adaptif yaklaşımların önemi vurgulanmıştır. Leaderboard'da ilk sırada yer alan 0.505 skoruna ulaşmak için daha gelişmiş stratejilere ihtiyaç olsa da, mevcut çalışmalar RNA 3D yapı tahmininde kullanılacak etkili yaklaşımlar hakkında değerli içgörüler sağlamıştır.