Int. J. Adv. Eng. Pure Sci. 2021, 33(4):592-608
DOI: 10.7240/jeps.920075

RESEARCH ARTICLE / ARAŞTIRMA MAKALESİ

Arayüz Mutasyonlarının Protein Etkileşimlerine Tesirini Tahmin Eden Algoritmalarla HADDOCK'un Performansının Karşılaştırılması

Predicting the Impact of Interfacial Mutations with HADDOCK: A Comperative Study

Mehdi KOŞACA[1,2], Eda ŞAMİLOĞLU[1,2], Ezgi KARACA[1,2]

[1]İzmir Biyotıp ve Genom Merkezi, Balçova, 35340, İzmir, Türkiye
[2]İzmir Uluslararası Biyotıp ve Genom Enstitüsü, Dokuz Eylül Üniversitesi, Balçova, 35340, İzmir, Türkiye

Öz
Hücresel süreçler proteinlerin birbirleriyle yaptıkları etkileşimlerinin üzerinden ilerler. Bilinen protein-protein etkileşimleri, etkileşim arayüzlerinde meydana gelen nokta mutasyonları ile yeniden düzenlenebilir. Bu düzenleme sonucunda, mevcut etkileşimler bozulabilir ve bu durum, kanser ve nörodejenaratif hastalıkların oluşmasına yol açabilir. Mutasyonların bu kadar hayati bir etkisinin olabilmesi, onların protein etkileşimleri üzerindeki tesirinin tahminini hesaplamalı biyolojinin aktif çalışma alanlarından biri haline getirmiştir. Mevcut mutasyon etki tahmin algoritmalarının yanında, ünlü kenetlenme programı HADDOCK, protein-protein etkileşim arayüzünde görülen mutasyonların, ayrıntılı bir şekilde modellenmesine olanak sağlamaktadır. Bu çalışmamızda, HADDOCK'un literatürde önerilen kullanım parametrelerini optimize ederek, mutasyon tahmin performansını iyileştirmeyi hedefledik. Bu kapsamda yaptığımız karşılaştırma çalışmamızda, HADDOCK'un en uygun parametre seçkisi ile bile alternatif bir kuvvet alanı temelli mutasyon tahmin algoritması olan EvoEF1'in performansını geçemediğini ortaya koyduk. Bunun yanında, EvoEF1'in performansını EvoEF2, FoldX ve UEP tahmin algoritmalarınınki ile karşılaştırdığımızda, EvoEF1'in en iyi performansı gösterdiğini gözlemledik. Dolayısıyla, bu çalışmamızın sonucu olarak, EvoEF1 programının kuvvet alanı temelli nokta mutasyonun etkisini tahmininde öncelikli olarak kullanılmasını önermekteyiz.
Anahtar Kelimeler: Nokta mutasyonu modellemesi, bağlanma afinitesi, HADDOCK, protein-protein etkileşimi, örnekleme optimizasyonu.

Abstract
Cellular processes are mediated by a diverse range of protein-protein interactions. Point mutations observed across protein-protein interfaces may lead to severe rearrangements of the available interaction patterns. Consequently, native interactions might be disrupted, leading to serious diseases, such as cancer. This has put predicting the impact of interfacial mutations as one of the central questions of the computational structural biology field. To this end, we probed the capabilities of the famous molecular docking program HADDOCK, in modelling and scoring of the interfacial mutations. For this, we traced the impact of number of mutations sampled on HADDOCK's prediction capacity. We observed that even for higher (more thorough) sampling numbers, HADDOCK could not exceed the performance of another force field-based algorithm EvoEF1, a program specifically tuned to predict the impact of interfacial mutations. EvoEF1 retained its leading status when we compared its performance to EvoEF2, FoldX and UEP algorithms. Thus, we propose EvoEF1 as a viable force field-based predictor for estimating the effect of point mutations on protein-protein interactions.
Keywords: Modeling point mutation, binding affinity, HADDOCK, protein-protein interaction, HADDOCK sampling optimization

I. GİRİŞ

Hücredeki yaşamsal süreçlerin devamlılığı proteinlerin birbirleriyle yaptığı etkileşimler aracılığıyla belirlenir [1]. Hücresel metabolizmada yer alan protein etkileşimlerinin karmaşık yapısına bakıldığında, canlılığın devamı için reaksiyonların geri dönüşümlü olması ve geçici etkileşimlerle kurulması gerektiği görülür [2]. Geçici etkileşimlerin oluştuğu bölge olan etkileşim arayüzü, ya da kısaca arayüz, tanımı ve kullanım amacına göre değişiklik gösterse de genellikle iki proteinin etkileştiği bölgede, 5-7Å'luk bir çap içinde kalan tüm etkileşimler olarak tanımlanır. Etkileşim arayüzündeki amino asitlerde meydana gelen nokta mutasyonları, bahsettiğimiz bu geçici etkileşimleri etkileyerek proteinin fonksiyonunu bozabilir [3]. Ayrıca, kansere, Alzheimer, Hungtington gibi nörodejeneratif hastalıklara ya da Mendeliyen hastalıklara ve nadir hastalıklara sebep olabilir [4]. Bu hastalıkların oluşum mekanizmalarının anlaşılması için, arayüzdeki amino asitleri hedefleyen protein mühendisliği yaklaşımları kullanılmaktadır. Bu yaklaşımlar, arayüz amino asitlerini nokta mutasyonları ile yeniden düzenleyerek, etkileşimi istenen yönde (zayıf-kuvvetli) kontrol edebilir [5]. Arayüz etkileşimlerinin incelenmesi, aynı zamanda evrimsel süreçte korunmuş ve günümüze ulaşmış doğal protein varyantlarının etkileşime etkisini anlamamız açısından önemlidir.

Sorumlu yazar: Ezgi KARACA, Tel: +90533 456 60 96, e-posta: ezgi.karaca@ibg.edu.tr
Gönderilme: 19.04.2021, **Düzenleme:** 10.11.2021, **Kabul:** 10.11.2021

Nokta mutasyonu seviyesindeki hassas değişiklikleri, tekrarlı olarak deneysel yöntemlerle çalışmak büyük bir emek ve pahalı deney düzenekleri gerektirmektedir. Bu sebeple, deneysel maliyeti ve zamanı azaltmak, arayüzde oluşturulan mutasyonların etkisini ölçmek ve değerlendirmek için pek çok hesaplamalı yöntem geliştirilmiştir. Bu yöntemler, etkileşen partnerlerin geometrik olarak birbirlerine ne kadar uyumlu oldukları, evrimsel dizi benzerlikleri, fiziksel enerji değerleri, istatistiksel olarak etkileşimde bulunma potansiyelleri gibi değişik bilgileri kullanarak, mutasyon sonucunda oluşacak yeni etkileşimlerin etkisini doğru bir şekilde tanımlamaya çalışır [6]. Mutasyon etkisini tahmin eden programların performansı genellikle SKEMPI isimli veri kümesi temel alınarak ölçülmektedir. SKEMPI veri kümesi, literatürden toplanmış 7086 nokta mutasyonunun, protein etkileşimlerinde yarattığı serbest bağlanma enerji değişim değerlerini (ΔΔG) içermektedir [7]. ΔG terimi bağlanma serbest enerjisi olmak üzere, ΔΔG terimi mutant ve yabanıl kompleksin bağlanma serbest enerjilerinin farkını temsil eder (Eşitlik 1). ΔΔG, mutasyonun proteinlerin bağlanması üzerindeki etkisinin bir ölçütüdür [8].

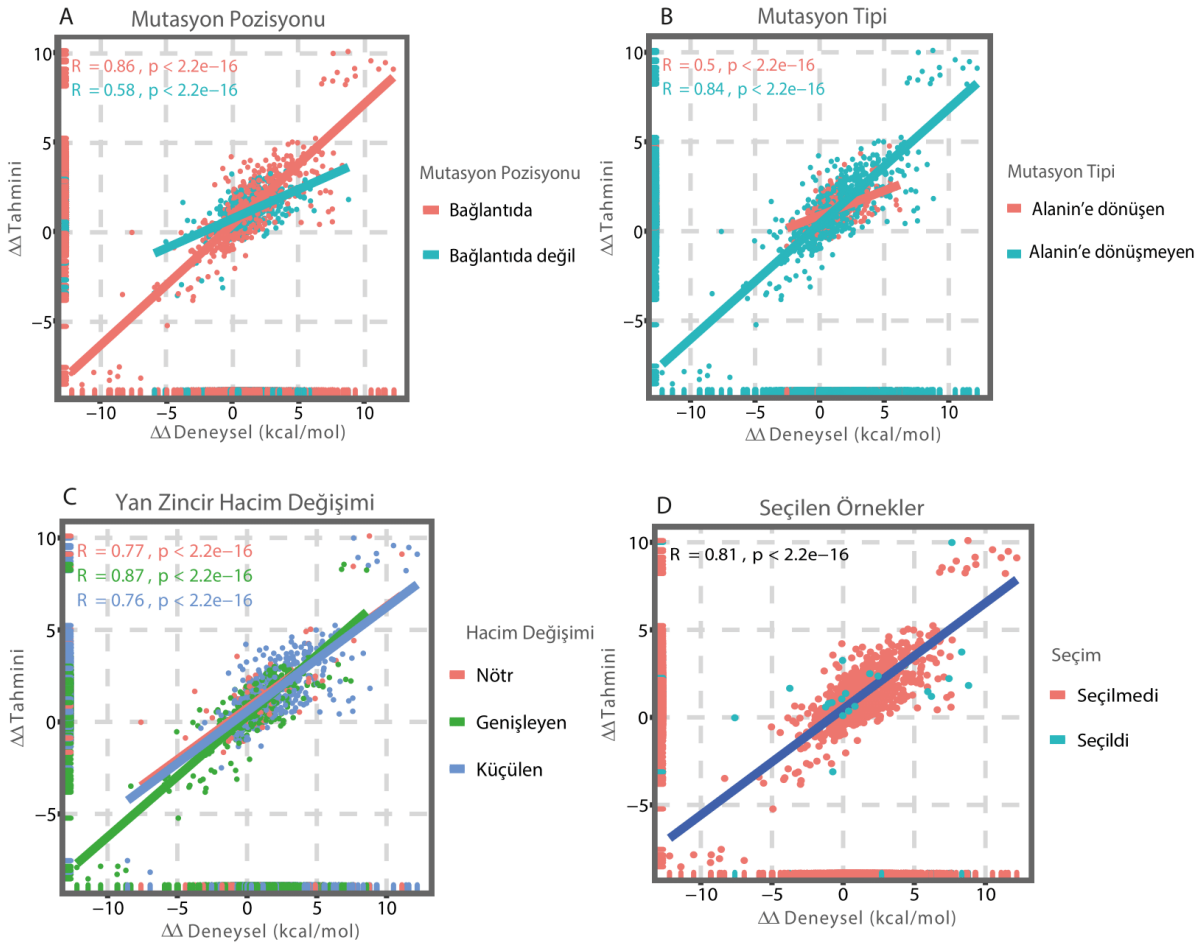
$$\Delta\Delta G_{mutasyon} = \Delta G_{mutant\text{-}kompleks} - \Delta G_{yabanıl\text{-}kompleks} \quad (1)$$

SKEMPI'nin daha odaklı ve bir alt kümesi olan DACUM [9], SKEMPI'den farklı olarak mutasyonlara dair enerji hesaplama yöntemi bilgilerini, mutasyonun etkileşimin neresinde olduğu (arayüzde ya da değil) ve mutasyon tipi (alanin ya da alanin harici diğer amino asitlere dönüşenler) gibi bilgileri içermektedir. Bu set üzerinde performansı denenmiş ve yaygın olarak kullanılan programlar arasında HADDOCK, EvoEF1 ve FoldX bulunmaktadır. Yakın zamanda yayınlanan, protein-protein komplekslerindeki mutasyonları sınıflandırmak için geliştirilmiş UEP algoritmasının makalesinde [10], EvoEF1 ve FoldX programlarının nokta mutasyon etkisini tahmin etmede en iyi performansı gösterdikleri Interactome3D veri kümesi [11] üzerinden ortaya konmuştur. Bu iki programın ortak özelliği, mutasyonu yapı üzerinde modellemeleri ve kuvvet alanı (*ing. force field*) temelli bir skorlama fonksiyonuna sahip olmalarıdır. FoldX [12], mutasyonların bağlanma enerjisinde yarattığı değişimi ölçmek için geliştirilmiştir. FoldX'in skorlama fonksiyonu, katsayıları deneysel verilerden elde edilmiş pek çok enerji teriminin ağırlıklı toplamlarından meydana gelir. EvoEF1 ise protein tasarımında bağlanmayı tanımlamak üzere optimize edilmiş fiziksel bir enerji fonksiyonudur. Bu fonksiyon; amino asit içi etkileşim, aynı ve farklı zincirdeki amino asitler arası etkileşim enerjilerinin ağırlıklı toplamlarından oluşur [13]. Bu iki yöntemin hızları karşılaştırıldığında EvoEF1'in FoldX'ten daha hızlı olduğu görülmüştür. Bu çalışmada HADDOCK'un performansı bahsi geçen karşılaştırmalı set üzerinde test edilmemiş olsa da iki yeni makalede HADDOCK'un, virüs ve konak protein varyantlarının COVID19 gelişimindeki etkilerini iyi derecede tahmin edebildiği ortaya konmuştur [14, 15]. Bu çalışmaların ilkinde insan Anjiyotensin dönüştürücü enzim 2 (hACE2) doğal varyantlarının etkileşim profilleri HADDOCK programı ile araştırılmıştır [15]. Diğerinde ise güncel verilerden yola çıkarak virüsün insanlardan hayvanlara geçişinde yapı bilgisinin, diziden daha geçerli olduğu HADDOCK yapı modellemeleri ile ortaya konmuştur [14]. Ancak bu veya benzer çalışmalarda, HADDOCK'un mutasyon etkisini modellemedeki başarısının örneklem (üretilen model sayısı: genel geçer durumda 20 ya da 50 olarak belirlenmiştir. bkz. Materyal ve Metod) parametrelerine ne kadar bağımlı olduğu araştırılmamıştır. Bu açıktan yola çıkarak HADDOCK örneklem parametrelerinin optimizasyonunu yapıp, bu yeni parametrelerin tahmin üzerindeki etkisini, yukarıda listelediğimiz diğer kuvvet alanı temelli algoritmalar ile karşılaştırarak ortaya koyduk. Çalışmamızın, nokta mutasyon etkilerini modellemek isteyen araştırmacılara önemli bir yol gösterici olacağını düşünmekteyiz.

II. MATERYAL VE METOD

2.1. Veri Kümesi Seçimi

Bu çalışmamızda SKEMPI'nin [7] bir alt kümesi olarak seçilmiş DACUM [9] veri kümesi kullanılmıştır (https://github.com/haddocking/DACUM). DACUM veri kümesindeki mutasyonlar, fizikokimyasal özelliklerine göre [*mutasyonun pozisyonu (bağlantıda olan (ing:loop), bağlantıda olmayan (ing:nonloop); mutant amino asit tipi (alanine dönüşen, alanin harici diğer aminoasitlere dönüşen); amino asit yan zincirinin hacim değişimi (genişleyen, küçülen ve nötr)]* sınıflandırılmıştır. Bu veri setinde ayrıca, bir makine öğrenmesi algoritması olan iSEE'nin mutasyon sonucunda gerçekleşen ΔΔG değişim tahminleri ile deneysel olarak ölçülmüş ΔΔG değerleri de bulunmaktadır [19]. iSEE yöntemi, HADDOCK skorlarını bir random forest algoritması dahilinde kullanarak, DACUM veri seti üzerinde HADDOCK'tan daha iyi bir performans göstermiştir. Biz de bu makaleyi HADDOCK'un performansı için üst sınır ve alt sınır tanımını yapacak verileri barındırdığı için test kümemizin seçiminde referans olarak kullandık. Bu doğrultuda, iSEE tarafından ΔΔG değerleri doğru tahmin edilen (10 adet) ve edilemeyen (10 adet) olmak üzere, 20 kompleks seçtik (Tablo 1). Bu seçimimizde doğru tahmin noktaları Şekil 1D'de regresyon çizgisine yakın, yanlış tahminler ise uzak olan verilerden seçtik (seçilen veriler Şekil 1D'de turkuaz renk ile işaretlenmiştir). Veri kümesinden bir alt küme oluşturulurken herhangi bir eğilim yaratmamaya dikkat edilmiş, örnekler tüm veri kümesini kapsayacak şekilde mümkün olduğunca her bir özellikten eşit sayıda olacak şekilde seçilmiştir (Şekil 1, Tablo 1).



Şekil 1. Seçilen örneklerin fizikokimyasal özelliklerine göre dağılımı.

Tablo 1. Alt küme olarak seçilen örneklerin PDB kodları ve mutasyon listesi. iSEE tarafından ΔΔG skoru doğru tahmin edilen örnekler (*) ile belirtilmektedir.

PDB Kodu	Zincir Kodu	Yabanıl Rezidü	Pozisyon	Mutant Rezidü
1A22*	A	ASP	26	ALA
1A4Y*	B	ARG	32	ALA
1CSE*	I	LEU	38	ASP
1E96*	A	ASN	26	HIS
1F47	A	LYS	14	ALA
1IAR*	A	THR	13	ALA
1KTZ*	B	SER	25	ALA
1MAH	A	TRP	276	ARG
1PPF	I	LEU	18	TRP
1R0R*	I	THR	12	GLN
1SIB	I	LYS	46	ARG
1XD3*	B	LYS	27	ARG
1Y33	I	PRO	39	THR
1Z7X	W	TYR	434	ALA
2FTL	I	LYS	15	SER
2G2U*	B	ARG	144	ALA
2PCC	A	GLU	290	ALA
3BTM	I	MET	13	LYS
3SGB*	I	GLY	26	HIS
3SGB	I	LEU	12	PRO

2.2. Verilerin Elde Edilmesi ve Hazırlanması

Oluşturulan alt kümenin PDB uzantılı dosyaları DACUM'dan indirilmiştir. Buradan indirilen yapılar pdb-tools (https://github.com/haddocking/pdb-tools)[16] python paketine ait **pdb_occ** (yer tutma değeri *(ing: occupancy)* sütununu 1.0'a ayarlamak için kullanıldı), **pdb_selaltloc** (çoklu konformasyona sahip amino asitlerde her atom için yer tutma değeri en yüksek olan atomu otomatik olarak seçip diğer atom konformasyonunu silmek için kullanıldı) ve **pdb_delhetatm** (PDB yapısındaki kimyasalları ve suları temizlemek için kullanıldı) araçları kullanılarak HADDOCK programının okuyabileceği şekilde düzenlenmiştir. UEP, FoldX, EvoEF1, EvoEF2 ve deneysel ΔΔG skorları UEP kütüphanesinden (https://github.com/pepamengual/UEP) [10], ΔHADDOCK skorları ise mutant HADDOCK skoru ile yabanıl tip HADDOCK skoru arasındaki fark alınarak elde edilmiştir. Kullanılan bütün algoritmalar ΔΔG skor tahminlerini Eşitlik (1)' deki formülü kullanarak yapmaktadır [8, 10, 12, 13, 17]. UEP'in hesaplama setinde bulunmayan örneklerin 3-boyutlu mutasyon modelleri, FoldX (4.0) ve EvoEF1 programlarının yönergeleri takip edilerek üretilmiştir. Diğer örneklerin FoldX ve EvoEF1 mutasyon modelleri, daha öncesinden UEP çalışması için modellenip yapıları depolanan UEP kütüphanesinden, HADDOCK mutasyon modelleri ise, sonuç çıktısı olarak gelen dosyada "structure>it1>water" yolu takip edilerek 50 ve 250 örneklem içerisinde en iyi HADDOCK skoruna sahip olan (en düşük değer) modeller alınarak elde edilmiştir.

2.3. Nokta Mutasyonu Uygulanması ve HADDOCK Guru Interface Parametreleri

Koordinat dosyasında mutasyon yapma işlemi, yabanıl amino asit üç harf kodunun mutant amino asit üç harf kodu ile değiştirilip, yan zincir atomlarının da (N, CA, C, O ve CB atomları haricindeki atomlar) silinmesi ile yapılmıştır. Eksik yan zincir atomları HADDOCK tarafından tamamlanmıştır. Örneklerin hem mutant hem de yabanıl tip formlarının HADDOCK skorları, Şekil S1'de belirtilen HADDOCK2.2 Guru Interface parametreleri (https://wenmr.science.uu.nl/)[18] ile hesaplanmıştır. HADDOCK, mutasyonun etkisi örneklemek için rastgele seçilmiş farklı hızlarla başlayan kısa moleküler dinamik simülasyonları uygular. Her bir örneklenen yapı, farklı simülasyondan gelen örneklemin sonucudur. HADDOCK kullanılarak yapılan arayüz mutasyon modellerinde, yaygın olarak 20 ve 50 örneklem sayıları kullanılmıştır [19].

2.4. Programların RMS değerlerinin Karşılaştırılması

Deneysel ve mutant protein yapıları arasındaki karşılaştırmalar, ortalama karekök sapması (RMS) değerinin hesaplanması ile yapılmıştır. RMS analizi HADDOCK, FoldX ve EvoEF1 ile üretilen modellerle yapılmıştır. EvoEF2'de mutasyon modeli üretebilmesine rağmen deneysel verilerle performansı daha kötü olduğu için karşılaştırmalara dahil edilmemiştir. Modellerin RMS karşılaştırmaları HADDOCK'un ilk 250 ve 50 örneklem içerisindeki en iyi skora sahip yapılar referans alınarak; "HADDOCK 250 - HADDOCK 50", "HADDOCK 250 - EvoEF1", "HADDOCK 250 - FoldX", ve "FoldX - EvoEF1" çiftleri arasında, PyMol (https://pymol.org/2/) programı kullanılarak yapılmıştır.

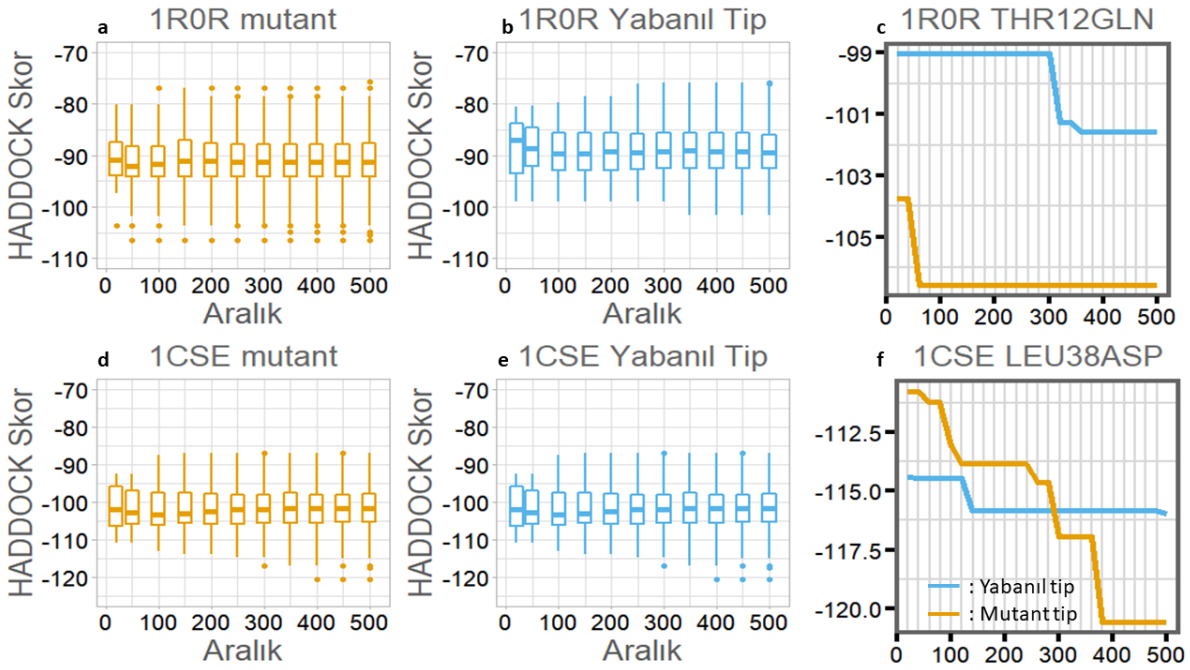
III. BULGULAR VE TARTIŞMA

3.1. HADDOCK Örnekleme Optimizasyonu

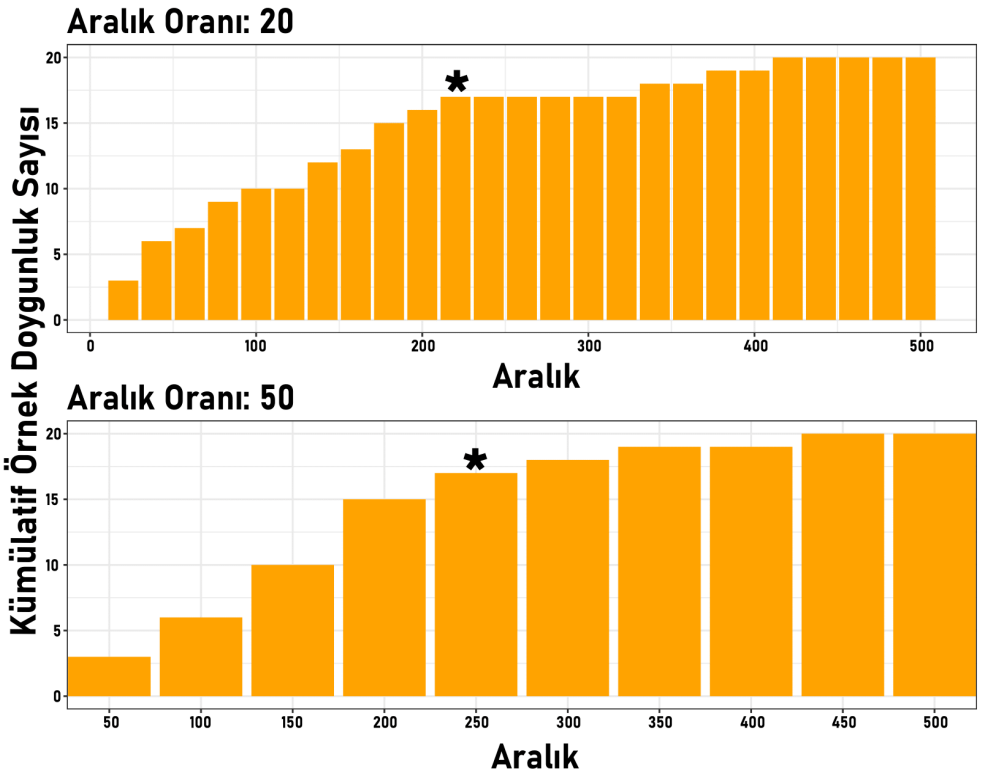
HADDOCK için literatürde önerilen örneklem sayıları 20 veya 50 olarak geçmektedir (bkz. Yöntem). Çalışmamızda, bu örneklem sayılarının yeterli olup olmadığı, Tablo 1'de verilen veri seti kullanılarak araştırılmıştır. Bunun için örneklem sayısında en yüksek sınır olarak 500 kullanılmıştır. 500 örneğin skor dağılımları, oluşum sırasına göre 20'lik, 50'lik ve 100'lük aralıklara ayrılarak kutu ve çizgi grafiği üzerinden incelenmiştir. Kutu grafiğinin içi, doğası gereği, genel geçer değerleri barındırdığından, skoru diğer modellere göre daha düşük olan minimum uç değerlere (*ing:outlier*) odaklanılmıştır [20]. Şekil S2'de seçilen 20 örneğin HADDOCK skorlarının 50'lik aralıklara göre kümülatif dağılımı, kutu ve çizgi grafikleri ile gösterilmiştir. Oluşturulan grafiklerin her aralığı incelenerek, en düşük enerjiye, en az örnekleme sayısıyla ulaşan aralık optimum örnekleme sayısı olarak belirlenmiştir. Aktarılan sürecin görselleştirilmesi adına, veri setinden alınan 1R0R ve 1CSE komplekslerinin örnekleme optimizasyonu Şekil 2'de gösterilmiştir.

1R0R örneğinin çizgi grafiğinde yabanıl tip için en iyi HADDOCK skoru ilk 350, mutant tip için ise ilk 50 yapı içerisinde elde etmiştir (Şekil 2c). Yabanıl tipte 300. aralığa kadar HADDOCK skoru değişmeyip, örneklenen 300. yapıdan 350. yapıya geçerken HADDOCK skorunda yaklaşık iki birimlik bir iyileşme görülmüştür (Şekil 2b). Mutant tipte ise örneklenen ilk 50 yapıdan sonra HADDOCK skoru hiç değişmeyip 500. aralığa kadar aynı sayıda kalmıştır (Şekil 2a). HADDOCK skorunda yaklaşık iki birimlik bir iyileşme elde etmek için yabanıl tip için fazladan 300 model üretmek yerine bu örnek için optimum örneklem sayısı "50" olarak belirlenmiştir. 1CSE örneğinin çizgi grafiğinde yabanıl tipte örneklenen ilk 100 yapıdan 150 yapıya geçerken HADDOCK skorunda yaklaşık 1 birim, 450. yapıdan 500. yapıya geçerken ise yaklaşık 0.5 birimlik bir iyileşme meydana gelmiştir (Şekil 2f). Mutant tipte ise örneklenen ilk 50 yapı ile 500. yapı arasında çok değişkenli bir şekilde geçişler meydana gelmiştir (Şekil 2c). 350. yapıdan 400. yapıya geçerken yaklaşık 2.5 birimlik bir iyileşme olduktan sonra HADDOCK skoru hiç değişime uğramadan 500. yapıya kadar aynı skor ile devam etmiştir. Yabanıl tip için fazladan 350 yapı, mutant tip için ise fazladan 200 yapı üretmek yerine bu örnek için optimum örneklem sayısı ise "300" olarak belirlenmiştir. Yukarıdaki örneklerde olduğu gibi, her protein kompleksi için optimum örneklem sayısına ulaşılan nokta, doygunluk

noktası (*ing: saturation point*) olarak tanımlanmıştır. Tüm veri seti üzerinde, farklı bölümleme sayıları için gözlemlenen doygunluk noktaları histogram grafiği ile analiz edilmiştir. Şekil 3'te görüldüğü gibi HADDOCK için örnekleme parametresini 250 model üretecek şekilde ayarlamak, veri seti üzerindeki çoğu model için en düşük enerjili yapıyı üretmeye yeterli olmuştur.



Şekil 2. 1R0R ve 1CSE örneklerinin yabanıl tip ve mutant formlarının minimum HADDOCK skorlarının kümülatif olarak kutu ve çizgi grafiklerlerle gösterimi.



Şekil 3. Seçilen 20 örneğin 20'lik (A) ve 50'lik (B) aralıklara göre kümülatif doygunluk oranının histogram grafikleri ile gösterimi. (*) optimum doygunluk miktarını belirtmektedir.

3.2. Programların ΔΔG Tahminlerinin Deneysel ΔΔG Skorlarıyla Karşılaştırılması

Mutasyonun bağlanmaya etkisini tahmin etmeye yarayan kuvvet alanı temelli HADDOCK, FoldX, EvoEF1 ve EvoEF2 programların performanslarının, deneysel verilerle karşılaştırılması Şekil 4'te gösterilmiştir. Bu programlar mutasyonları bir kuvvet alanı etkisinde modeller ve enerji değişimlerini hesaplar. Bu programların yanında UEP, sisteme yüklenen yapıdan çok hızlı ve basit bir şekilde ama herhangi bir yapı üretmeden, en az iki başka atomla daha etkileşim yapma kriterlerine uyan arayüzdeki amino asit mutasyonlarının bağlanmaya olan etkilerini hesaplar. Kullandığımız veri seti içerisinde UEP kriterlerine uymayan mutasyonlar Tablo 2.'de gösterilmiştir.

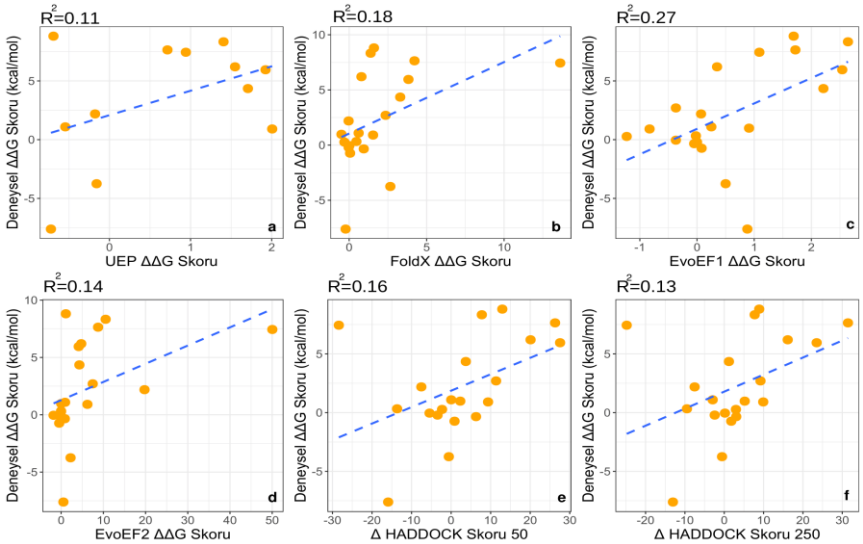
Tablo 2. UEP hesaplama kriterlerine uymayan alt küme verileri

PDB Kodu	Zincir Kodu	Yabanıl Rezidü	Pozisyon	Mutant Rezidü
1A22	A	ASP	26	ALA
1F47	A	LYS	14	ALA
1IAR	A	THR	13	ALA
1KTZ	B	SER	25	ALA
1SIB	I	LYS	46	ARG
1XD3	B	LYS	27	ARG
2G2U	B	ARG	144	ALA
3SGB	I	GLY	26	HIS

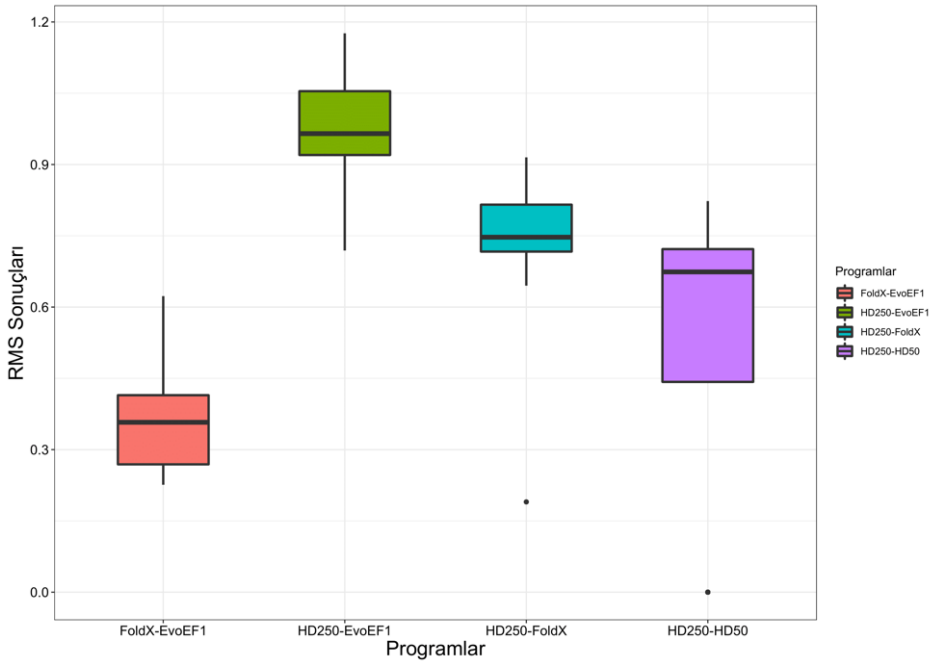
HADDOCK örnekleme optimizasyonunun ΔΔG tahmin performansını gözlemlemek için örneklerin klasik HADDOCK düzeneğinde kullanılan 50 örneklem sayısındaki minimum değerler de dikkate alınmıştır. Deneysel verilerle en iyi korelasyonu veren program EvoEF1 (n=20'de R^2=0.27), ardından da FoldX'tir (n=20'de R^2=0.18). EvoEF2, EvoEF1'e kıyasla deneysel verilerle daha kötü korelasyon göstermiştir. Bu farklılık EvoEF1 ve EvoEF2 arasındaki monomerik amino asit - amino asit etkileşimlerindeki kuvvet etkisinden kaynaklanmaktadır [17]. EvoEF1'de van der Waals çekim kuvvetinin EvoEF skoruna etkisi, itme kuvvetinin etkisinden fazlayken EvoEF2'de bu durum tam tersidir. Van der Waals itme kuvvetinin fazlalığı, genişleyen yan zincir mutasyon tiplerinde sterik çarpışmalar için daha yüksek bir ceza vererek ΔΔGStabilite ve ΔΔGBağlanma tahminlerinde, EvoEF2'nin EvoEF1'den daha kötü performans göstermesine yol açmıştır [17]. Yapmış olduğumuz HADDOCK örnekleme optimizasyonu sonucunda, 250 örneklem sayısının, HADDOCK örnekleme fonksiyonunu iyileştirmesine rağmen istatistiksel olarak deneysel ΔΔG skorlarıyla anlamlı bir korelasyona götürmediği gözlenmiştir.

3.3. Programların Mutant Yapı Modellerinin Karşılaştırılması

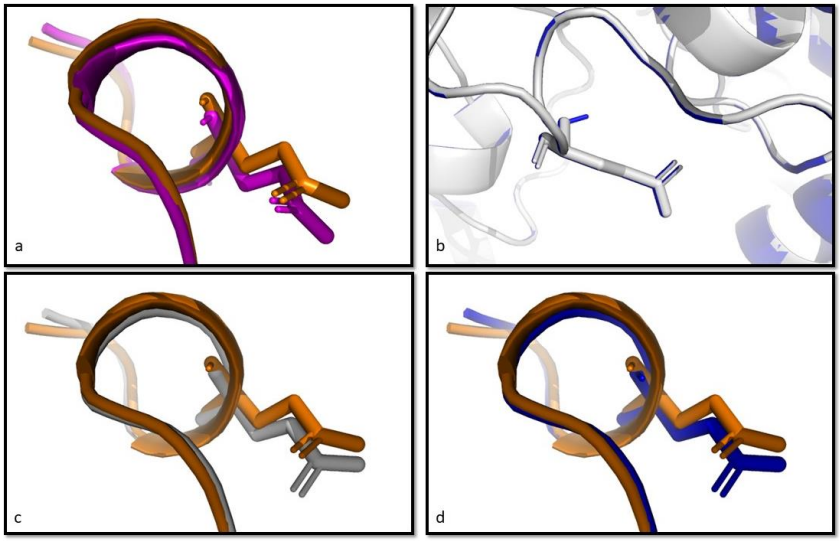
Programların yapı modelleme performanslarını karşılaştırmak için üretilen yapılar arasında RMS değerleri hesaplanmıştır (Şekil 5). Küçük RMSD değeri, yapıların birbirine daha çok benzediği anlamına gelmektedir [19]. Yapı modeli üretebilen programlar (HADDOCK, EvoEF1 ve FoldX) arasında yapılan ikili model benzerliği karşılaştırmasında birbirine en çok benzeyen modellerin FoldX ve EvoEF1 (RMSD ortalaması = 0.37 Å) tarafından üretildiği görülmüştür. HADDOCK 250 ve 50 örneklem sayılı modeller ise FoldX - EvoEF ikilisinden sonra birbirine en çok benzeyen modelleri üretmiştir (RMSD ortalaması = 0.49 Å). Şekil 6d'de 2PCC örneği üzerinden gösterilen model benzerliğinde de FoldX-EvoEF1 modellerinin benzerliği görülmektedir. HADDOCK - FoldX ve HADDOCK - EvoEF1 modellerinin RMSD ortalamaları sırasıyla 0.74 Å ve 0.93 Å olmuştur. Üretilen yapı modelleri arasında en büyük farklılığın HADDOCK 250 ve EvoEF1 modelleri arasında olduğu görülmüştür (Şekil 5). Bu bulgu, kullanılan yöntemlerin arasındaki performans farkında farklı mutasyon modelleme yaklaşımlarının da etkili olduğunu göstermektedir.



Şekil 4. Programların tahmin skorlarının deneysel ΔΔG bağlanma afiniteleri (kcal/mol) ile karşılaştırılması.



Şekil 5. FoldX-EvoEF1 (kırmızı), HADDOCK 250-EvoEF1 (yeşil), HADDOCK 250-FoldX (mavi) ve HADDOCK 250-HADDOCK 50 (mor) yapı ikililerinin RMS sonuçlarının kutu grafiği ile dağılımı (Å).



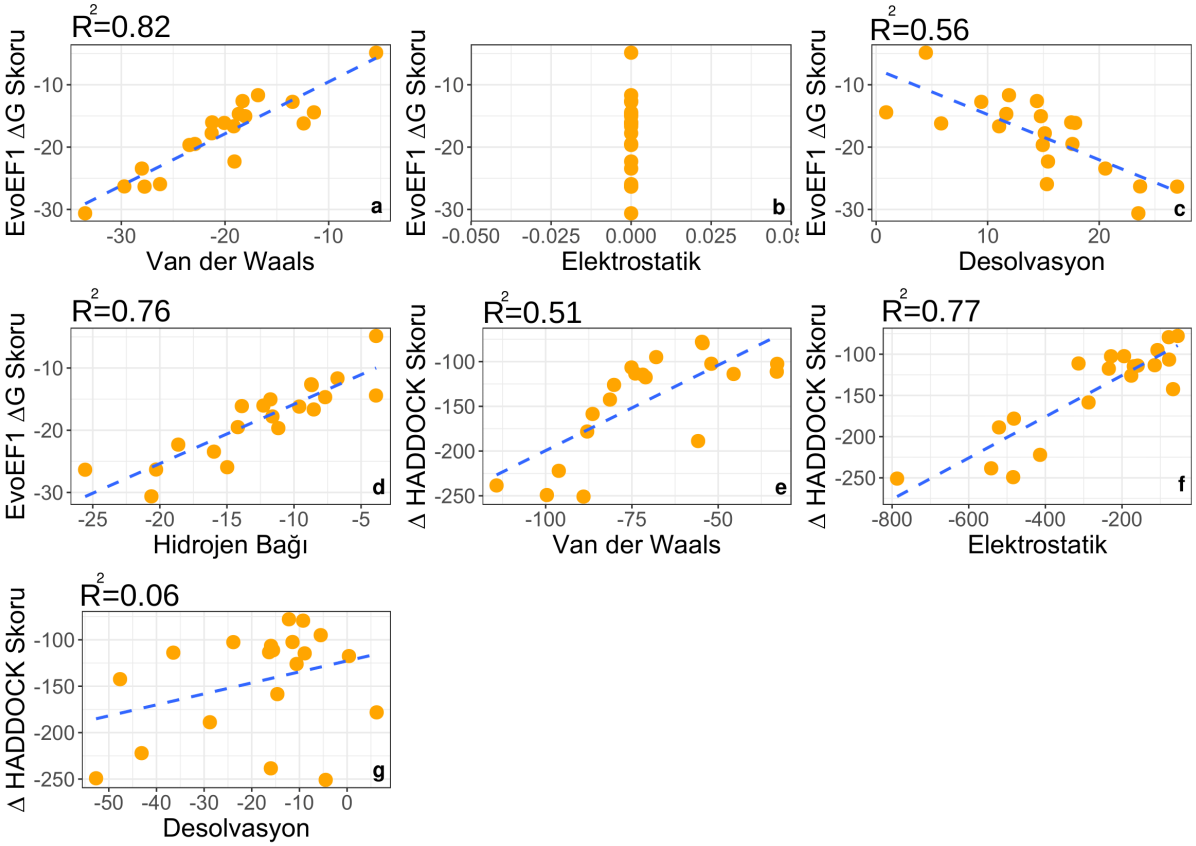
Şekil 6. 2PCC EA290A mutant yapısının farklı programlarca üretilmiş modellerinin RMS karşılaştırmaları (PyMol).Turuncu: HADDOCK 250, mor: HADDOCK 50, gri: FoldX, lacivert: EvoEF1 yapılarını belirtmektedir

Farklı skorlama fonksiyonlarının modelleme başarısı üzerindeki etkisini tespit etmek için HADDOCK ve EvoEF1 programlarının skorlama fonksiyonları araştırılmıştır. Bu programların skorlama formülleri sırasıyla Eşitlik (2) ve (3)'te gösterilmektedir [13, 18].

$$E_{HADDOCK} = 1.0 * E_{VDW} + 0.2 * E_{ELEK} + 1.0 * E_{DESOLV} + 0.1 * E_{AIR} \quad (2)$$

$$E_{EvoEF} = E_{VDW} + E_{ELEK} + E_{HB} + E_{DESOLV} - E_{REF} \quad (3)$$

E_{VDW}, E_{ELEK}, E_{HB}, E_{DESOLV}, E_{AIR}, E_{REF} sırasıyla toplam van der Waals, elektrostatik (Coloumb Kuvveti), hidrojen bağlanma, desolvasyon, belirsiz etkileşim kısıtlaması (*ing:Ambiguous Interaction Restraints*) ve protein sekansının referans enerjisini belirtir. Veri kümemiz içerisindeki 20 örneğin HADDOCK ve EvoEF1 programlarıyla üretilen yapı modellerinde etkileşim arayüzünü oluştururken baskın olarak kullandıkları kuvvetleri belirlemek için her enerji teriminin skora olan katkısı korelasyon analizi ile belirlenmiştir (Şekil 7). Buna göre HADDOCK için seçilen 20 örneğin arayüz etkileşimlerinde elektrostatik (R^2 = 0.77) ve van der Waals (R^2 = 0.51) enerji terimlerinin etkisi fazlayken desolvasyon enerjisinin (R^2 = 0.05) etkisi diğer iki terime kıyasla daha azdır (Şekil 7 a, b, c). EvoEF1'de ise sırasıyla van der Waals (R^2 = 0.81), hidrojen bağlanma (R^2 = 0.75) ve desolvasyon (R^2 = 0.55) enerji terimleri arayüz etkileşimlerinde baskındır (Şekil 7 d, e, f, g).



Şekil 7. HADDOCK ve EvoEF1 skorlarının enerji terimleriyle korelasyonu.

HADDOCK, yapı hesaplama aracı olarak CNS kuvvet alanını [21] kullanmaktadır. Moleküller arası ve moleküler içi enerjileri, OPLS kuvvet alanının bağlı olmayan (*ing:nonbonded*) parametreleri [22] kullanılarak 8.5Å sınırı (*ing:cut-off*) dahilindeki tüm van der Waals ve elektrostatik enerjileri değerlendirerek hesaplamaktadır [23]. HADDOCK'tan farklı olarak EvoEF1 hesaplama verimliliği için maksimum etkileşim üst sınır değerini 6Å olarak ayarlamıştır. Ayrıca EvoEF1, van der Waals enerjisini hesaplamak için CHARMM19 kuvvet alanının [25] bir parçası olan Lennard-Jones 12-6 potansiyelinin modifiye edilmiş halini kullanır. EvoEF1 elektrostatik için ise kısmi yüklü atomların elektrostatik etkileşimlerini PARSE metodunu kullanır [24]. Bu yöntem, yüklü iki atom arasındaki mesafeye göre hesaplama yapmaktadır. Bu mesafenin ötesindeki kısmi yüklü etkileşimler için elektrostatik değeri "0" kabul edilir [13]. Van der Waals kuvvet etkisinin EvoEF1'de HADDOCK'a kıyasla daha etkili olmasının sebebi belirlenen farklı kuvvet alanının etkisi olabilir. Daha önce yapılan çalışmalarda CNS kuvvet alanını kullanan HADDOCK bağlanma afinitesi tahmini içinde kullanılmış fakat genel fonksiyonlara sahip olduğu için anlamlı istatistiksel sonuçlara ulaşılamamıştır [26]. EvoEF1 için seçilen kuvvet alanları ve belirlenen sınır değerlerinin bağlanma afinitesi tahmin performansında HADDOCK'a kıyasla daha doğru sonuçlar vermesinin sebebi olabileceği düşünülmektedir.

IV. SONUÇ VE ÖNERİLER

Protein-protein etkileşimlerinde nokta mutasyonlarının etkisini tahmin etmeye yarayan birçok program farklı enerji terimleri ve kriterler kullanılarak geliştirilmiştir. Yapılan bu çalışmada HADDOCK'un protein-protein etkileşimlerinde nokta mutasyonunun etkisini tahmin performansı, literatürdeki diğer kuvvet alanı temelli bağlanma afinitesi tahmin programları olan EvoEF1, EvoEF2, ve FoldX ile karşılaştırılarak analiz edilmiştir. Deneysel verilerle en iyi korelasyonu veren programların sırasıyla EvoEF1, FoldX, EvoEF2, HADDOCK olduğu Şekil 4'te yapılan ΔΔG hesaplama performansı analizi sonucunda belirlenmiştir. HADDOCK için yapılan optimizasyon işlemi sonucunda örneklem sayısı arttırılarak daha düşük enerjili modeller üretilmesine rağmen deneysel verilerle anlamlı bir korelasyon elde edilememiştir. Kullanılan programlardaki bu performans farklılığının yapı modellemeye ne kadar bağlı olduğu RMSD analizi ile belirlenmiştir. Üretilen yapı modellerindeki en büyük farklılığın HADDOCK ve EvoEF1 ikilisi arasında olduğu bulunmuştur. Skorlama fonksiyonlarının performans etkisini gözlemlemek için HADDOCK ve EvoEF1 programlarının kullanmış oldukları kuvvet alanları ve belirledikleri cut-off değerleri incelenerek seçilen örneklerin arayüz etkileşimlerindeki baskın kuvvetleri belirlenmek istenmiştir. Yapılan analizlerde EvoEF1 için seçilen kuvvet alanı ve belirlenen cut-off değerlerinin, bağlanma afinitesini tahmin performansında

HADDOCK'a kıyasla daha anlamlı sonuçlar vermekte etkisi olabileceği görülmüştür. Ayrıca EvoEF1'in çok hızlı mutasyon modellemesi yapması ve nokta mutasyonu etkisini hızlı tahmin etmesi ile HADDOCK ve FoldX'ten ayrılmaktadır. Tüm bu sonuçlar dahilinde, EvoEF1'in nokta mutasyonu etkisini kuvvet alanı temelli olarak tahmin etmek için kullanılabilecek en iyi program olduğu tespit edilmiştir. Bu programı kullanarak ileride hızlı bir şekilde protein-protein etkileşimleri üzerinde çalışmalar yapılabilir, yeni arayüz profili hazırlanarak bağlanmayı iyileştirebilecek ya da kötüleştirebilecek nokta mutasyonları hızlı bir şekilde belirlenebilir.

TEŞEKKÜRLER

2019-TA-02 Çağrı kodlu ve 3393 proje numaralı bu çalışma, Türkiye Sağlık Enstitüleri Başkanlığı (TÜSEB) tarafından desteklenmiştir. Desteklerinden ötürü TÜSEB'e ve Bilim Akademisi Genç Bilim İnsanları Ödül Programları'na (BAGEP) teşekkür ederiz. Ayrıca yardımlarından ve yol göstericiliğinden dolayı Mehmet Ergüven'e, makalenin kritik okumasını ve düzenlemelerini yaptığı için Ayşe Berçin Barlas'a, Büşra Savaş'a ve Burcu Özden'e teşekkür ederiz.

KAYNAKLAR

[1] Stites, W. (1997). Protein−Protein Interactions: Interface Structure, Binding Thermodynamics, and Mutational Analysis. Chemical Reviews, 97(5), 1233-1250. https://doi.org/10.1021/cr960387h

[2] Hein, M. Y., Hubner, N. C., Poser, I., Cox, J., Nagaraj, N., Toyoda, Y., Gak, I. A., Weisswange, I., Mansfeld, J., Buchholz, F., Hyman, A. A., & Mann, M. (2015). A human interactome in three quantitative dimensions organized by stoichiometries and abundances. Cell, 163(3), 712–723. https://doi.org/10.1016/j.cell.2015.09.053

[3] Subramanian, S., & Kumar, S. (2006). Evolutionary anatomies of positions and types of disease-associated and neutral amino acid mutations in the human genome. *BMC genomics*, *7*, 306. https://doi.org/10.1186/1471-2164-7-306

[4] Gonzalez, M. W., & Kann, M. G. (2012). Chapter 4: Protein interactions and disease. PLoS computational biology, 8(12), e1002819. https://doi.org/10.1371/journal.pcbi.1002819

[5] Krohl, P. J., Ludwig, S. D., & Spangler, J. B. (2019). Emerging technologies in protein interface engineering for biomedical applications. Current opinion in biotechnology, 60, 82–88. https://doi.org/10.1016/j.copbio.2019.01.017

[6] Karaca, E., & Bonvin, A. M. (2013). Advances in integrative modeling of biomolecular complexes. Methods (San Diego, Calif.), 59(3), 372–381. https://doi.org/10.1016/j.ymeth.2012.12.004

[7] Jankauskaite, J., Jiménez-García, B., Dapkunas, J., Fernández-Recio, J., & Moal, I. H. (2019). SKEMPI 2.0: an updated benchmark of changes in protein-protein binding energy, kinetics and thermodynamics upon mutation. Bioinformatics (Oxford, England), 35(3), 462–469. https://doi.org/10.1093/bioinformatics/bty635

[8] Geng, C., Xue, L., Roel-Touris, J. and Bonvin, A. (2021). Finding the ΔΔ G spot: Are predictors of binding affinity changes upon mutations in protein–protein interactions ready for it? WIREs Computational Molecular Science, 2019. 9(5). https://doi.org/10.1002/wcms.1410

[9] Geng, C., Vangone, A., & Bonvin, A. (2016). Exploring the interplay between experimental methods and the performance of predictors of binding affinity change upon mutations in protein complexes. Protein engineering, design & selection : PEDS, 29(8), 291–299. https://doi.org/10.1093/protein/gzw020

[10] Amengual-Rigo, P., Fernández-Recio, J., & Guallar, V. (2020). UEP: an open-source and fast classifier for predicting the impact of mutations in protein-protein complexes. Bioinformatics (Oxford, England), btaa708. Advance online publication. https://doi.org/10.1093/bioinformatics/btaa708

[11] Mosca, R., Céol, A., & Aloy, P. (2013). Interactome3D: adding structural details to protein networks. Nature methods, 10(1), 47–53. https://doi.org/10.1038/nmeth.2289

[12] Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., & Serrano, L. (2005). The FoldX web server: an online force field. Nucleic acids research, 33(Web Server issue), W382–W388. https://doi.org/10.1093/nar/gki387

[13] Pearce, R., Huang, X., Setiawan, D., & Zhang, Y. (2019). EvoDesign: Designing Protein-Protein Binding Interactions Using Evolutionary Interface Profiles in Conjunction with an Optimized Physical Energy Function. Journal of molecular biology, 431(13), 2467–2476. https://doi.org/10.1016/j.jmb.2019.02.028

[14] Rodrigues, J., Barrera-Vilarmau, S., M C Teixeira, J., Sorokina, M., Seckel, E., Kastritis, P. L., & Levitt, M. (2020). Insights on cross-species transmission of SARS-CoV-2 from structural modeling. PLoS computational biology, 16(12), e1008449. https://doi.org/10.1371/journal.pcbi.1008449

[15] Sorokina, M., M C Teixeira, J., Barrera-Vilarmau, S., Paschke, R., Papasotiriou, I., Rodrigues, J., & Kastritis, P. L. (2020). Structural models of human ACE2 variants with SARS-CoV-2 Spike protein for structure-based drug design. Scientific data, 7(1), 309. https://doi.org/10.1038/s41597-020-00652-6

[16] Rodrigues, J., Teixeira, J., Trellet, M., & Bonvin, A. (2018). pdb-tools: a swiss army knife for molecular structures. F1000Research, 7, 1961. https://doi.org/10.12688/f1000research.17456.1

[17] Huang, X., Pearce, R., & Zhang, Y. (2020). EvoEF2: accurate and fast energy function for computational protein design. Bioinformatics (Oxford, England), 36(4), 1135–1142. https://doi.org/10.1093/bioinformatics/btz740

[18] van Zundert, G., Rodrigues, J., Trellet, M., Schmitz, C., Kastritis, P. L., Karaca, E., Melquiond, A., van Dijk, M., de Vries, S. J., & Bonvin, A. (2016). The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. Journal of molecular biology, 428(4), 720–725. https://doi.org/10.1016/j.jmb.2015.09.01489

[19] Geng, C., Vangone, A., Folkers, G. E., Xue, L. C., & Bonvin, A. (2019). iSEE: Interface structure, evolution, and energy-based machine learning predictor of binding affinity changes upon mutations. Proteins, 87(2), 110–119. https://doi.org/10.1002/prot.25630

[20] Karaca, E., Rodrigues, J., Graziadei, A., Bonvin, A., & Carlomagno, T. (2017). M3: an integrative framework for structure determination of molecular machines. Nature methods, 14(9), 897–902. https://doi.org/10.1038/nmeth.4392

[21] Brünger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J. S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T., & Warren, G. L. (1998). Crystallography & NMR system: A new software suite for macromolecular structure determination. Acta crystallographica. Section D, Biological crystallography, 54(Pt 5), 905–921. https://doi.org/10.1107/s0907444998003254

[22] Jorgensen, W. L., & Tirado-Rives, J. (1988). The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. Journal of the American Chemical Society, 110(6), 1657–1666. https://doi.org/10.1021/ja00214a001

[23] Dominguez, C., Boelens, R., & Bonvin, A. M. (2003). HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. Journal of the American Chemical Society, 125(7), 1731–1737. https://doi.org/10.1021/ja026939x

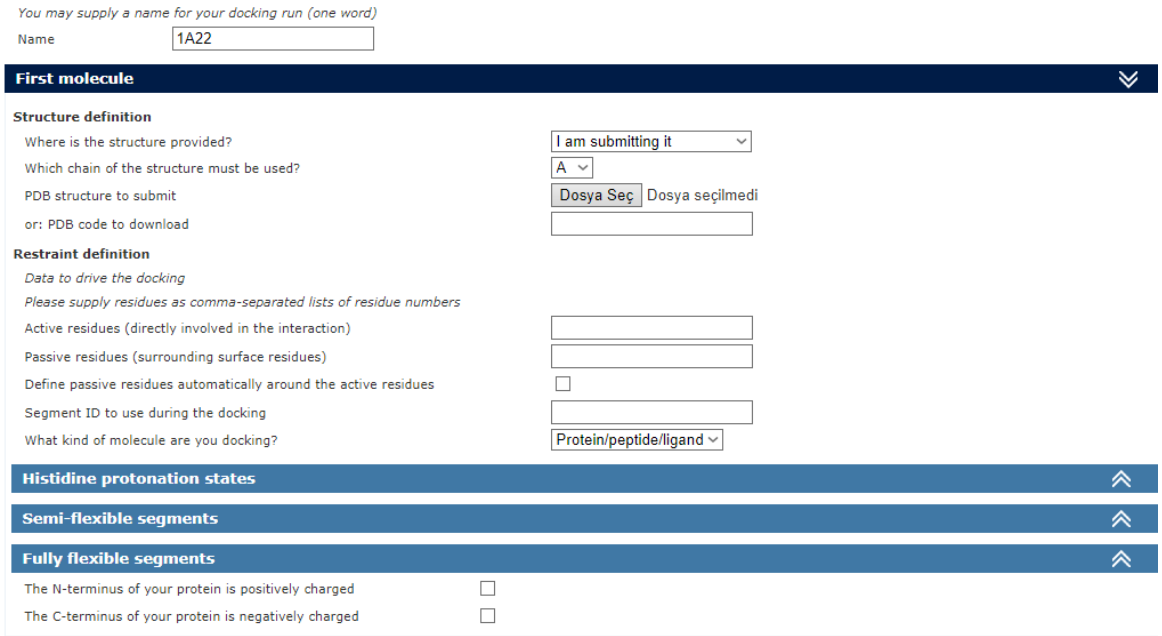
[24] Sitkoff, D., Sharp, K., & Honig, B. (1994). Accurate Calculation of Hydration Free Energies Using Macroscopic Solvent Models. The Journal Of Physical Chemistry, 98(7), 1978-1988. https://doi.org/10.1021/j100058a043

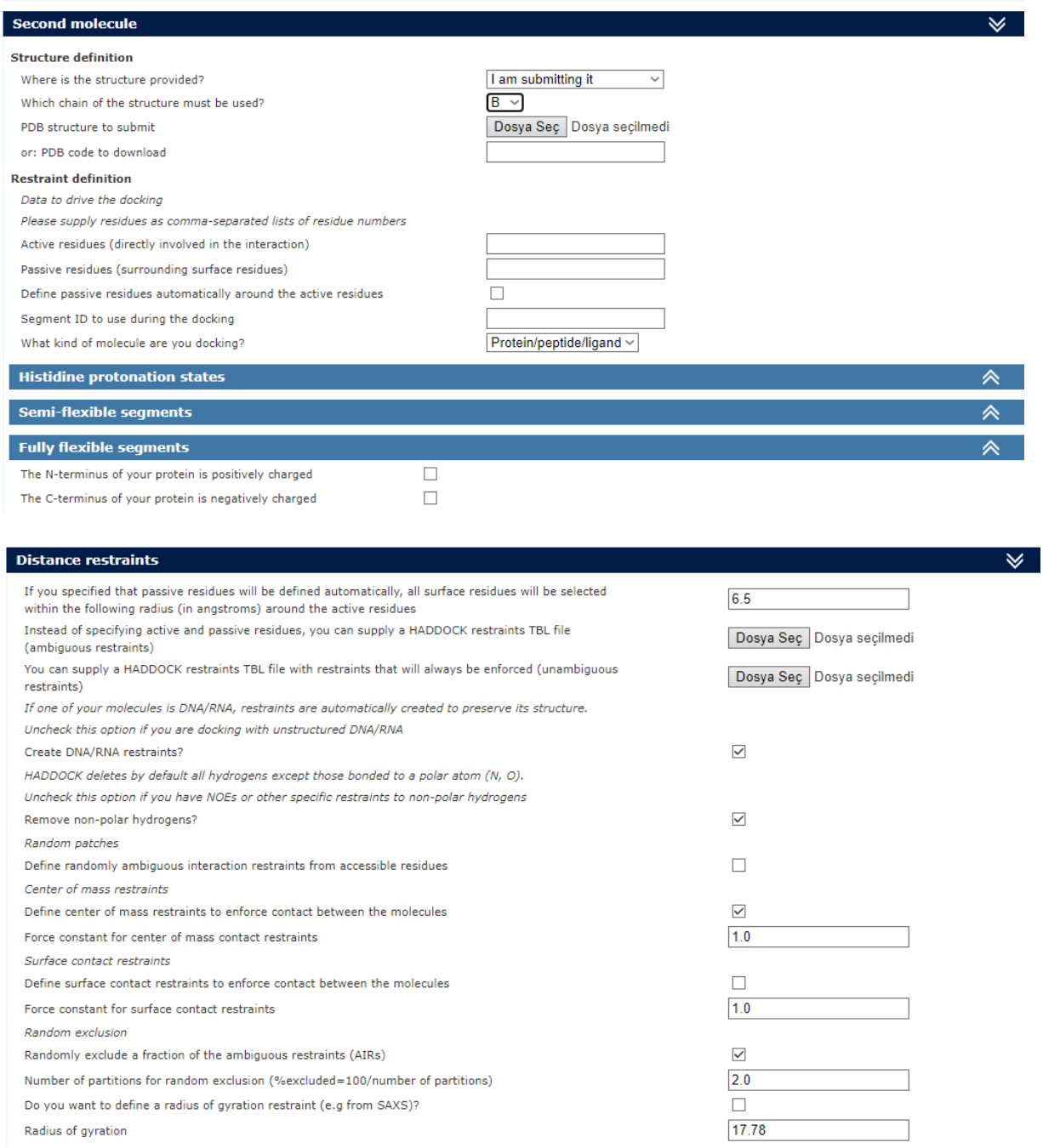
[25] Brooks, B. R., Brooks, C. L., 3rd, Mackerell, A. D., Jr, Nilsson, L., Petrella, R. J., Roux, B., Won, Y., Archontis, G., Bartels, C., Boresch, S., Caflisch, A., Caves, L., Cui, Q., Dinner, A. R., Feig, M., Fischer, S., Gao, J., Hodoscek, M., Im, W., Kuczera, K., … Karplus, M. (2009). CHARMM: the biomolecular simulation program. Journal of computational chemistry, 30(10), 1545–1614. https://doi.org/10.1002/jcc.21287

[26] Kastritis, P. L., & Bonvin, A. M. (2012). On the binding affinity of macromolecular interactions: daring to ask why proteins interact. Journal of the Royal Society, Interface, 10(79), 20120835. https://doi.org/10.1098/rsif.2012.0835

EK MATERYALLER

HADDOCK web sunucusunun, arayüz mutasyonu uygulamak için kullanımı:




Second molecule
Structure definition
Where is the structure provided?
I am submitting it
Which chain of the structure must be used?
B
PDB structure to submit
Dosya Seç
Dosya seçilmedi
or: PDB code to download
Restraint definition
Data to drive the docking
Please supply residues as comma-separated lists of residue numbers
Active residues (directly involved in the interaction)
Passive residues (surrounding surface residues)
Define passive residues automatically around the active residues
Segment ID to use during the docking
What kind of molecule are you docking?
Protein/peptide/ligand
Histidine protonation states
Semi-flexible segments
Fully flexible segments
The N-terminus of your protein is positively charged
The C-terminus of your protein is negatively charged
Distance restraints
If you specified that passive residues will be defined automatically, all surface residues will be selected within the following radius (in angstroms) around the active residues
6.5
Instead of specifying active and passive residues, you can supply a HADDOCK restraints TBL file (ambiguous restraints)
Dosya Seç
Dosya seçilmedi
You can supply a HADDOCK restraints TBL file with restraints that will always be enforced (unambiguous restraints)
Dosya Seç
Dosya seçilmedi
If one of your molecules is DNA/RNA, restraints are automatically created to preserve its structure.
Uncheck this option if you are docking with unstructured DNA/RNA
Create DNA/RNA restraints?
HADDOCK deletes by default all hydrogens except those bonded to a polar atom (N, O).
Uncheck this option if you have NOEs or other specific restraints to non-polar hydrogens
Remove non-polar hydrogens?
Random patches
Define randomly ambiguous interaction restraints from accessible residues
Center of mass restraints
Define center of mass restraints to enforce contact between the molecules
Force constant for center of mass contact restraints
1.0
Surface contact restraints
Define surface contact restraints to enforce contact between the molecules
Force constant for surface contact restraints
1.0
Random exclusion
Randomly exclude a fraction of the ambiguous restraints (AIRs)
Number of partitions for random exclusion (%excluded=100/number of partitions)
2.0
Do you want to define a radius of gyration restraint (e.g from SAXS)?
Radius of gyration
17.78

Sampling parameters

Number of structures for rigid body docking	500
Number of trials for rigid body minimisation	5
Sample 180 degrees rotated solutions during rigid body EM	☑
Number of structures for semi-flexible refinement	500
Sample 180 degrees rotated solutions during semi-flexible SA	☐
Solvent to use for the last iteration	water
Number of structures for the explicit solvent refinement	500
Epsilon constant for the electrostatic energy term	
Note that for explicit solvent refinement cdie with epsilon=1 is used	
Epsilon	10.0
Solvated docking mode	
Perform solvated docking	☐

Advanced sampling parameters

Do you want to cross-dock all combinations in the ensembles of starting structures?	
Turn off this option if you only want to dock structure 1 of ensemble A	
to structure 1 of ensemble B, structure 2 to structure 2, etc.	
Perform cross-docking	☑
Enable this option to multiply the number of structures in all iterations by the number of starting structure combinations.	
The number of combinations depends on the cross-docking parameter.	
If cross-docking is disabled, the number of combinations is the size of the first ensemble.	
If cross-docking is enabled, the number of combinations is the sizes of all ensembles multiplied.	
Multiply the number of calculated structures by all combinations	☐
Randomize starting orientations	☐
Perform initial rigid body minimisation	☐
Allow translation in rigid body minimisation	☐
initial seed for random number generator	917
it1 parameters	
temperature for rigid body high temperature TAD	2000
initial temperature for rigid body first TAD cooling step	2000
final temperature after first cooling step	500
initial temperature for second TAD cooling step with flexible side-chain at the inferface	1000
final temperature after second cooling step	50
initial temperature for third TAD cooling step with fully flexible interface	1000
final temperature after third cooling step	50
time step	0.002
factor for timestep in TAD	8
number of MD steps for rigid body high temperature TAD	0
number of MD steps during first rigid body cooling stage	0
number of MD steps during second cooling stage with flexible side-chains at interface	0
number of MD steps during third cooling stage with fully flexible interface	0
final solvated refinement	
number of steps for heating phase (100, 200, 300K)	100
number of steps for 300K phase	1250
number of steps for cooling phase (300, 200, 100K)	500
calculate explicit desolvation energy (note this will double the cpu requirements)	☐

Analysis parameters

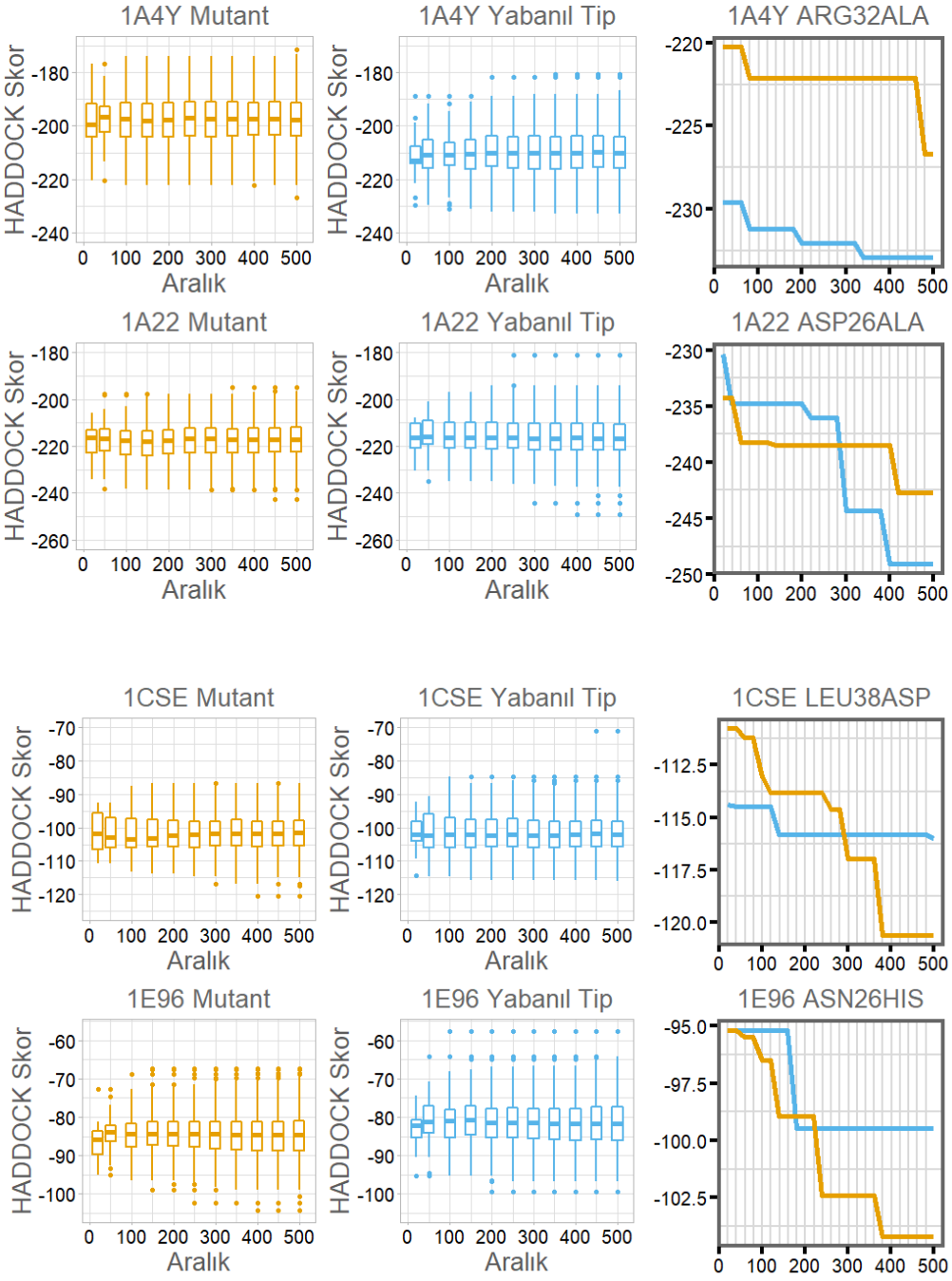
Number of structures to analyze	500
Cutoff distance (proton-acceptor) to define an hydrogen bond	2.5
Cutoff distance (carbon-carbon) to define an hydrophobic contact	3.9
After the final solvent refinement, write additional PDB files including solvent	☐

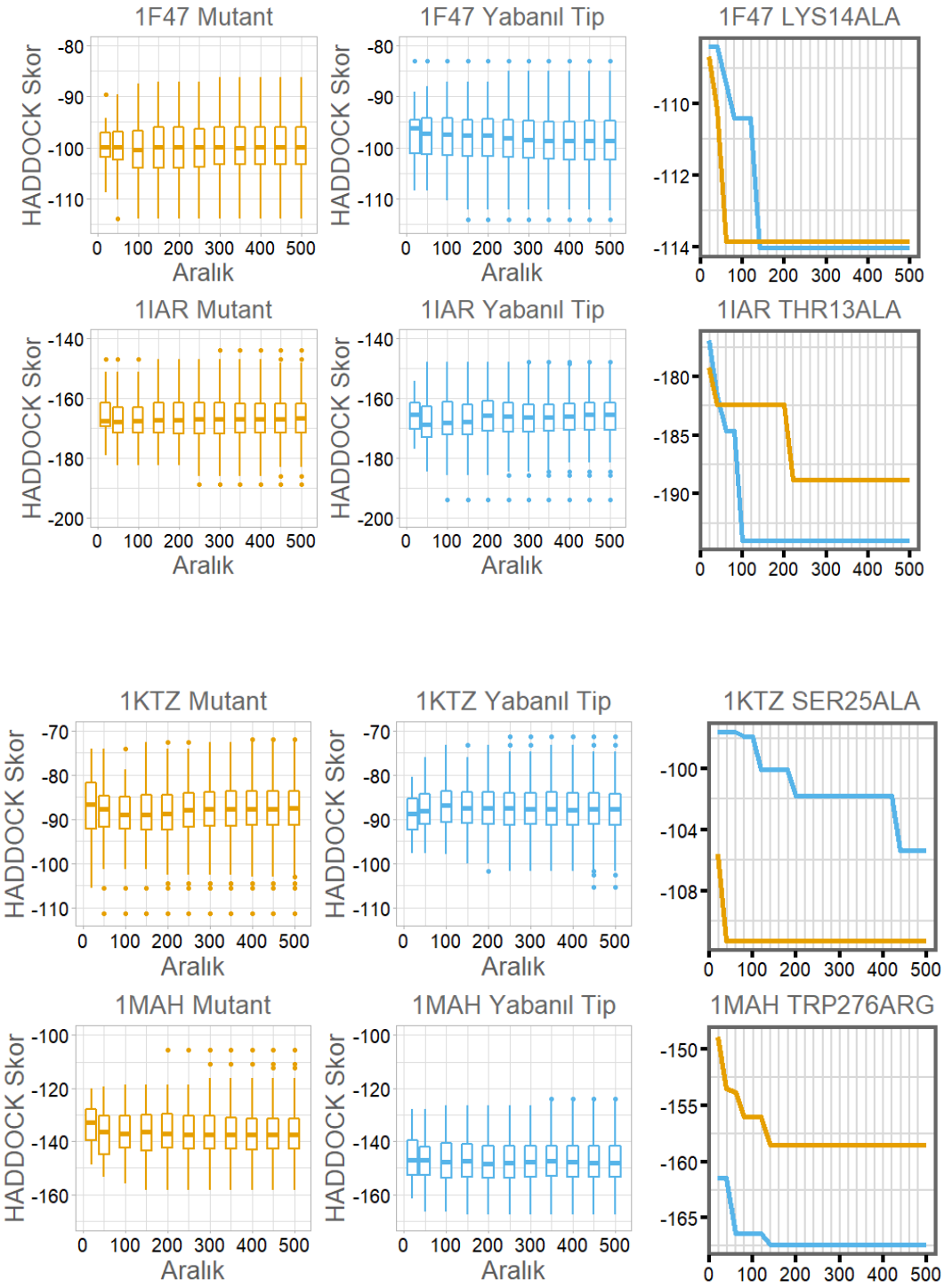
Username and password or EGI Check-in *(You should be registered before)*

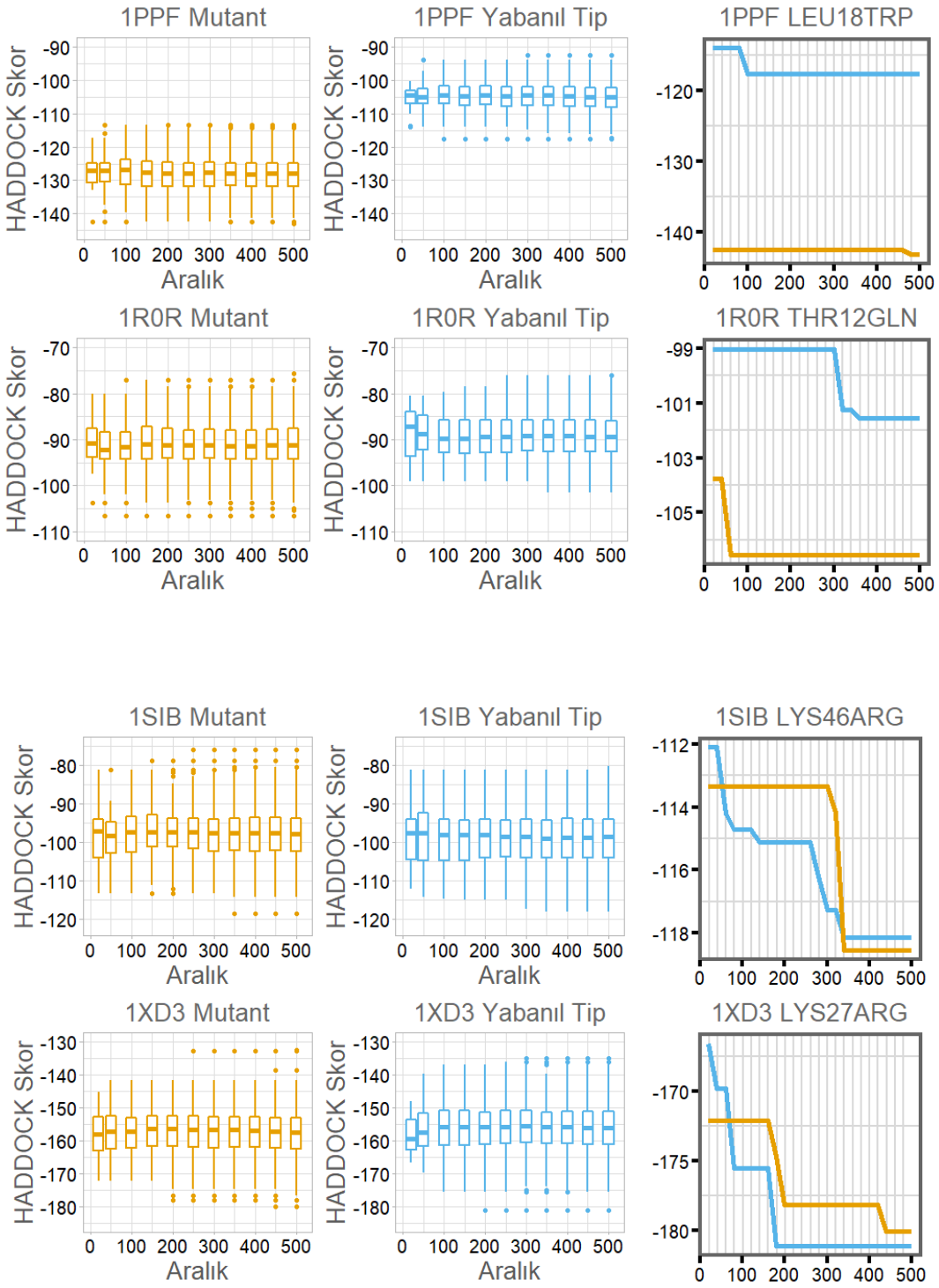
Is this a COVID-19 related submission ?	NO
Username	
Password	

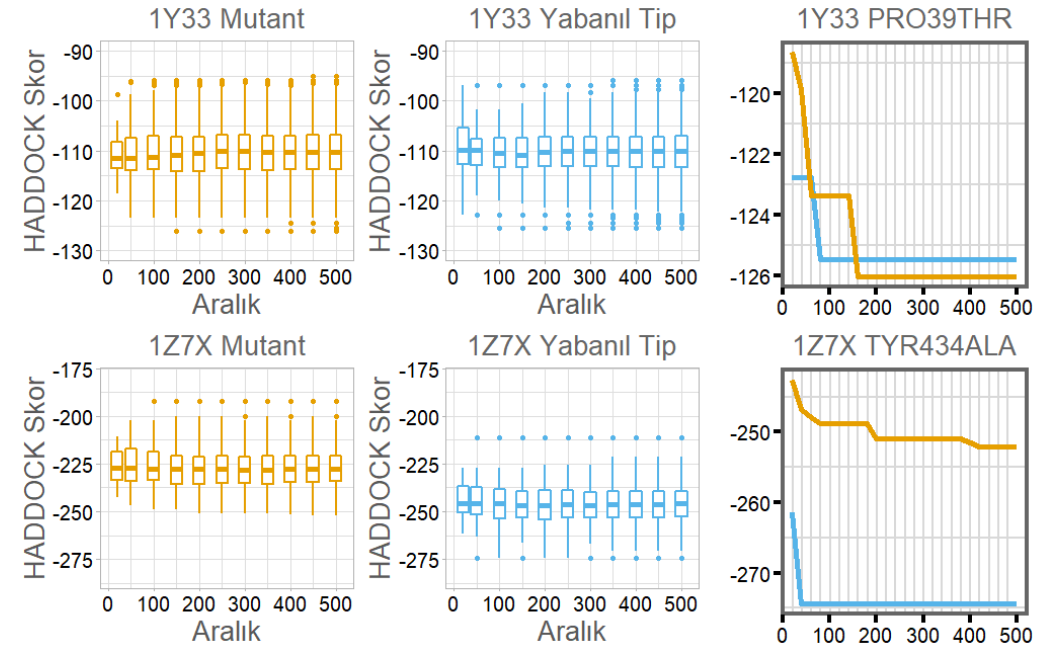
Gönder

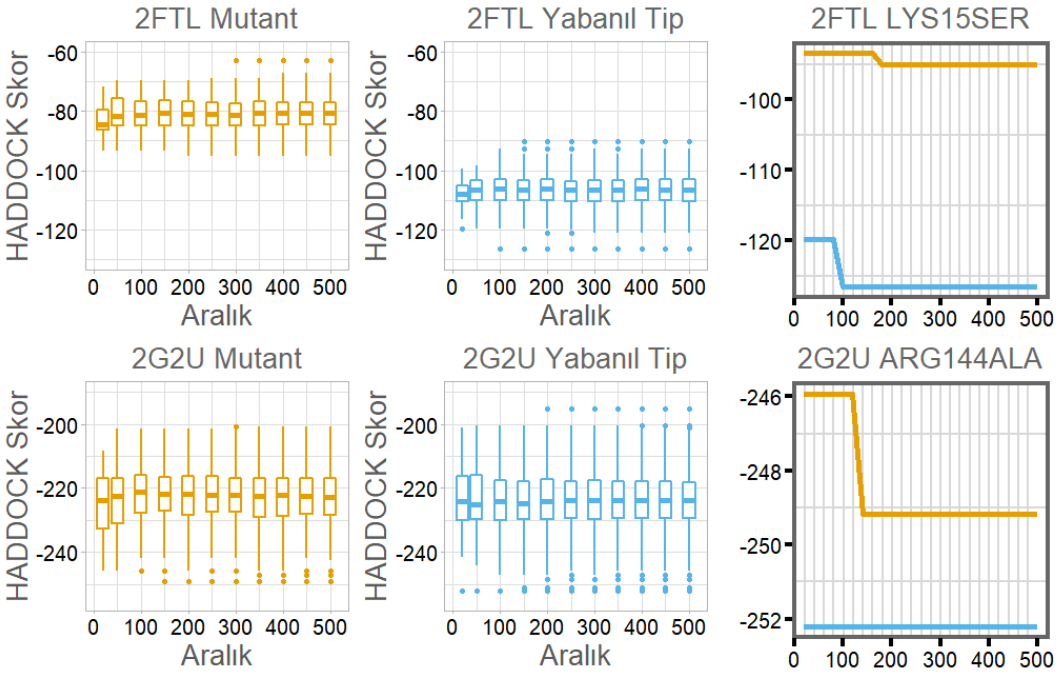
HADDOCK örneklem optimizasyonu için farklı komplekslerde enerji değişim hesabı:

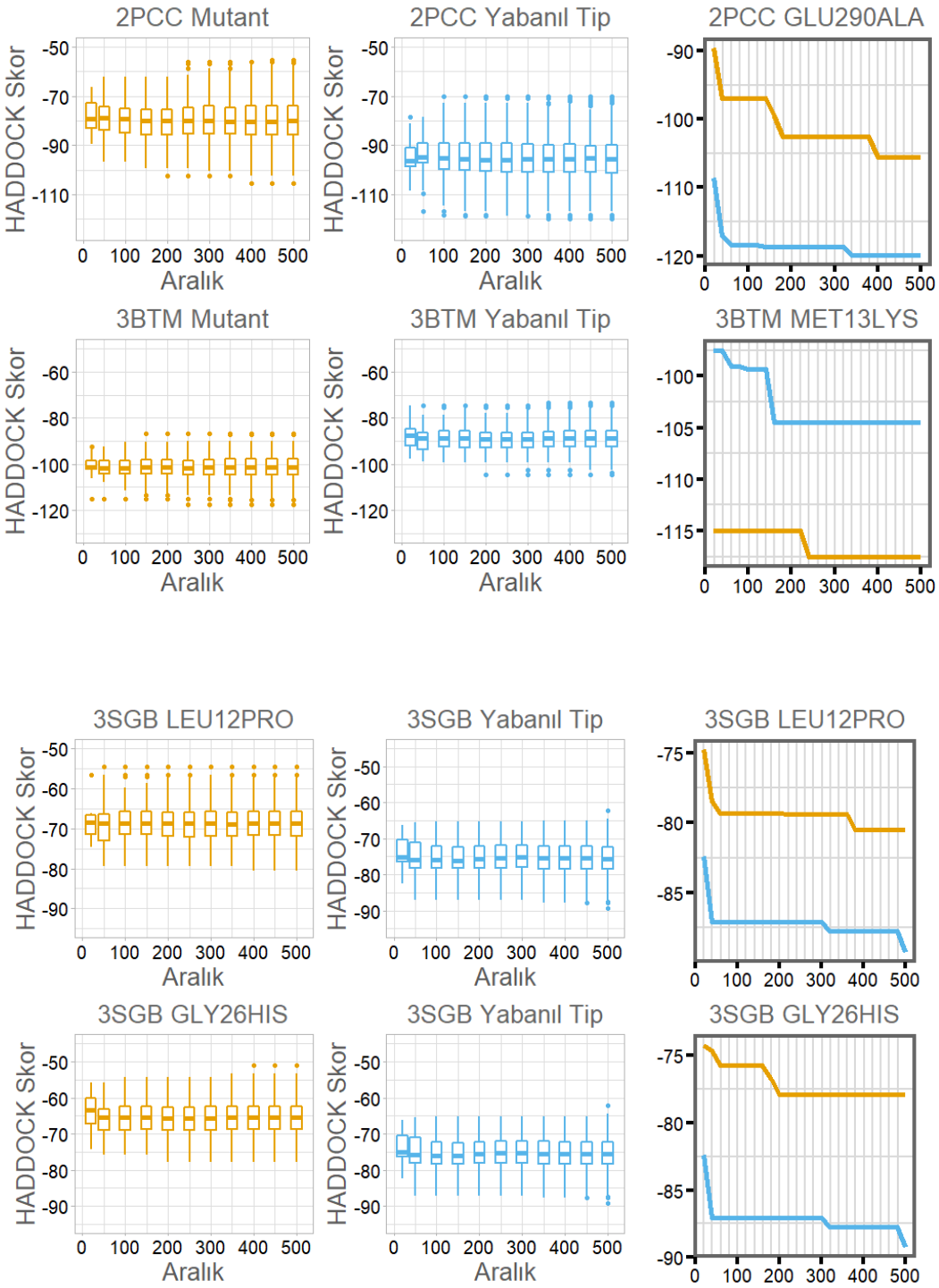



1F47 Mutant
1F47 Yabanıl Tip
1F47 LYS14ALA
1IAR Mutant
1IAR Yabanıl Tip
1IAR THR13ALA
1KTZ Mutant
1KTZ Yabanıl Tip
1KTZ SER25ALA
1MAH Mutant
1MAH Yabanıl Tip
1MAH TRP276ARG
HADDOCK Skor
Aralık


1PPF Mutant
1PPF Yabanıl Tip
1PPF LEU18TRP
1R0R Mutant
1R0R Yabanıl Tip
1R0R THR12GLN
1SIB Mutant
1SIB Yabanıl Tip
1SIB LYS46ARG
1XD3 Mutant
1XD3 Yabanıl Tip
1XD3 LYS27ARG
HADDOCK Skor
Aralık


1Y33 Mutant
1Y33 Yabanıl Tip
1Y33 PRO39THR
1Z7X Mutant
1Z7X Yabanıl Tip
1Z7X TYR434ALA
HADDOCK Skor
Aralık


2FTL Mutant
2FTL Yabanıl Tip
2FTL LYS15SER
2G2U Mutant
2G2U Yabanıl Tip
2G2U ARG144ALA
HADDOCK Skor
Aralık



Şekil S2. 50'lik aralıklara göre seçilen örneklerin HADDOCK skorlarının kümülatif dağılımlarının kutu ve çizgi grafiklerle gösterimi.