자유 에너지를 고려한 고리 구조 예측 방법을 적용한 단백질 구조 모델 정밀화

강범창,[1] 이규리,[1] 석차옥[1]
[1]서울대학교 화학부대한민국, 서울, 관악구, 관악로 1, 157-742
E-mail: bisulike@snu.ac.kr

단백질의 구조를 예측하기 위해서 구조가 알려져 있는 단백질 중 진화적으로 유사한 단백질의 구조 정보를 이용하는 Template Based Modeling (TBM) 방법이 현재까지 가장 효과적으로 많이 사용되고 있다. 단백질의 삼차 구조를 이루는 단위 중에서도 고리 부분은 효소 활성 부위 또는 리간드 결합 부위 등으로 작용하여 단백질의 생물학적 기능에 연관되어 있다. 하지만 진화적으로 가까운 단백질이어도 고리 부분은 서열이 잘 보존되지 않아 충분한 구조 정보를 주지 못하고 TBM 방법으로 고리 구조까지 정확히 예측을 할 수 없다. 따라서 TBM 방법으로 예측한 구조의 고리 부분을 주형 정보에 의존하지 않고 다시 예측하여 전체 구조를 정밀화하는 과정이 중요하다. 이번 연구에서는 이를 위해 자유 에너지를 고려한 고리 구조 예측 방법을 적용하여 그 효과를 검증해보았다.

Key Words: 주형 기반 단백질 구조 예측, 단백질 구조 모델 정밀화, 단백질 고리 구조 예측, 엔트로피, 자유 에너지, 콜로니 에너지

서론

단백질의 고리 구조를 예측한다는 것은 고리 부분을 제외한 단백질의 나머지 삼차 구조와 고리 부분의 서열이 알려져 있을 때 고리 부분의 삼차 구조를 예측하는 것을 말한다. 단백질을 구성하는 구조 단위를 기반으로 생각해볼 때 단백질 고리 구조는 알파 나선 구조나 베타 병풍 같은 이차 구조 단위들을 연결하는 부분이라고 정의할 수 있다. 한편, 진화적인 관점에서 생각해볼 때 고리 부분은 대개 서열이 보존이 안되며 따라서 단백질의 기능적 특이성을 지니게 하는 역할을 하는 부분이라고 정의할 수도 있다. 이렇게 단백질 기능에 관여하는 고리 부분은 단백질 접힘 과정에서 안전성에 관여하거나 단백질과 단백질의 상호작용, 리간드와의 결합, 효소 활성 등에도 연관이 된다..[1-4]

이처럼 단백질 고리 구조에 대한 이해가 단백질 기능에 직접적인 연관이 있어 중요하지만 단백질 구조를 예측하는 과정 중에서도 고리 구조를 예측하는 것은 어려운 문제이다. 현재 단백질 구조를 예측하는 방법 중 널리 쓰이는 방법으로는 구조가 알려진 단백질 중 진화적으로 유사한 단백질의 구조를 주형으로 삼아 예측하는 주형 기반 모델링 (Template based modeling, TBM) 방법이 있다. 하지만 앞에서 언급했던 것과 같이 단백질 고리 부분은 서열 상 보존이 잘 안되어있기 때문에 진화적으로 가까운 단백질의 구조를 주형으로 이용한다고 하여도 이러한 고리 부분의 구조 정보를 얻기에는 한계가 있게 된다.[5] 따라서 주형 구조에 의존하지 않는 *Ab initio*고리 구조 예측 방법을 추가적으로 적용 해야 할 필요가 있다. 단백질 구조 예측을 위한*Ab initio*접근법은 이용할 수 있는 주형 구조 없이 가능한 구조들

을 sampling 하고 에너지를 기반으로 한 scoring 을 통해 구조를 선택하는 기본적인 방식을 토대로 한다. 이 때 scoring 을 위해 대개 물리 화학 기반의 에너지 함수를 사용하는데, 대부분의 경우 entropy효과는 고려되지 않는다. 단백질의 고리 구조와 같은 경우 다른 이차 구조 단위보다 유연할 가능성이 높아 특히 엔트로피 효과를 추가적으로 고려해야 하는 것이 필수적이다. 따라서 이번 연구에서는 기존에 Ab initio 고리 구조 예측 방법으로 개발된 GalaxyFill[9]프로그램을 이용하여 sampling 된 고리 구조들을 평가할 때 엔트로피 효과까지 고려하고자 하여 colony energy[1] 개념을 도입하였다. 특히 실용적인 문제에 대한 검증을 하기 위해 주형 기반 모델링 방법인 GalaxyTBM[12] 을 이용해 실제 예측된 단백질 구조 모델의 고리 부분을 다시 예측해 정밀화하고자 했다.

이론 및 계산방법

단백질 구조 예측 대상 set 구성

국제적인 단백질 구조 예측 대회인 Critical Assessment of protein Structure Prediction (CASP) 에서 문제로 제출되었던 단백질 서열들을 대상으로 연구를 수행하였다. 최근에 제출되었던 문제 중 비교적 진화적으로 가까운 주형 단백질이 존재하는 단백질 서열을 선택적으로 16개 골랐다. 이는 실용적으로 생각했을 때 단백질 고리 구조 예측을 적용하여 구조 모델이 정밀화될 수 있는 지 여부를 반영하여 선택한 결과이다. 진화적으로 가까운 주형 단백질이 존재하지 않는 경우 단백질 고리 구조 예측을 적용해도 구조 모델을 정밀화하는 효과가 미미하기 때문이다. 예측된 단백질 구조 모델에서 추가적으로 정밀화해야 하는 고리 부분은 단백질의 밝혀진 실험 구조를 아는 상태에서 주형 기반으로 예측된 구조의 부정확한 특정 부분을 지정하였다. 이는 이번 연구의 대상을 예측해야 하는 단백질의 서열 정보만을 통해 추가적 정밀화가 필요한 고리 부분을 예측하는 또 다른 문제와 구분하기 위해서이다. 최종적으로 16개의 예측된 단백질 구조 모델에서 38 개의 고리 부분을 추가적으로 예측해 정밀화하였다.

단백질 구조 예측

비교적 진화적으로 가까운 주형 단백질 구조가 존재하는 16개의 단백질 서열에 대하여 에디슨에 탑재된 주형 기반 단백질 모델링 프로그램인GalaxyTBM[12]을 사용하여 예측된 단백질 구조 모델을 만들었다. 그 후 실험을 통해 알려져 있는 실험 구조와 예측된 구조 모델을 비교하여 특히 부정확한 고리 부분을 지정해 총 38개의 고리 부분을 선택하였다. 이 때 각 고리 부분마다 고리 부분을 둘러싼 주변 단백질 구조 모델의 정확도를 측정할 수 있다. 단백질 고리 부분으로부터 10 Å 내에 원자가 존재하는 주변 아미노산을 고리 부분의 환경이라 정의하고 그 환경에 대해 실험 구조와의 RMSD 를 계산하였다. 이 때 정밀화해야 하는 단백질 구조 모델의 정확도가 어느 한계 이하이면, 또는 고리의 환경의 정확도가 떨어지면 고리 구조의 예측 정확도도 매우 낮아지게 된다. 따라서 환경의 RMSD 가 3 Å 이상인 고리 부분은 제외하고 남은 31개 고리 부분에 대하여 추가적인 예측을 수행하였다.

단백질 구조 모델을 정밀화하기 위한 고리 구조 sampling

각각의 선별된 고리 부분에 대하여, 에디슨에 탑재된 단백질 고리 구조 예측 프로그램인 GalaxyFill[9]을 이용하여 2000개의 후보 구조들을 만들었다. GalaxyFill[9]은 알려진 실험 구조들

의 부분 조각 구조들을 조합하는 fragment assembly와 고리 닫힘 알고리즘인 analytic loop closure라는 방법을 이용하여 구조를 생성한다.

GalaxyFill 프로그램 자체는 가능한 고리 부분의 구조를 생성만 하기 때문에 어떤 구조가 더 적합한 지는 추가적인 scoring 을 통해 평가해야 한다. 이 때 더 정확한 에너지 계산을 위해서 예측된 구조를 한 번 더 최적화하기 위해 단백질 곁가지 최적화 프로그램인 GalaxySC[10, 11]를 사용하였다. 이 역시 에디슨에 탑재되어있는 프로그램이다. 그 뒤에GALAXY energy[13]로 각 고리 구조의 에너지를 계산한 후 물리적으로 비정상적인 구조를 제외시키기 위해 에너지가 낮은 500개 만을 다음 계산 단계를 위해 선택하였다.

자유 에너지를 고려한 고리 구조 scoring

이번 연구에서는 단백질 구조 모델을 정밀화하기 위한 추가적인 고리 구조 예측을 적용할 경우 자유에너지를 추가적으로 고려하는 것이 얼마나 도움이 되는 지 검증을 하고자 했다. 자유에너지를 고려하는 한 가지 방법으로 colony energy 라는 개념을 도입하였다. Wang[1]등이 기존 연구에서 제안한 것처럼, N개의 구조중 i번째 구조의 Colony Energy, $\Delta CE(i)$ 는 다음과 같은 형태를 갖는다.

$$\Delta CE(i) = -k_B T \ln\left[\sum_{j=1}^{N} \alpha(i,j) e^{-E_j / k_B T}\right], \quad (1)$$

일종의 확률 함수인 $\alpha(i,j) = \exp\left(-\mathrm{RMSD}_{ij}^{n} / \gamma\right)$ 는 i 구조의 자유에너지를 고려할 때 함께 ensemble 을 이루는 다른 j구조들의 기여도를 나타내는weight factor로,j구조와 i구조 사이의 구조 차이, 즉 Root Mean Square Deviation (RMSD) 를 이용해 정의된다. 즉, i 구조의 에너지는 i 와 구조적으로 가까운 j 구조들의 에너지에 영향을 받게 되며 구조적으로 가까운 j 구조들이 많을 경우 엔트로피가 증가하며 자유에너지는 낮아지는 것을 기술하게 된다. 자유에너지를 고려하기 전 위치에너지를 계산할 때 이용한 GALAXY energy의 크기 단위는 분자역학 기반의 CHARMM force field energy의 크기 단위의 열 배이기 때문에 Ej값은Galaxy energy를 10으로 나눈 값을 사용했다.[14] Colony energy 계산을 위해 필요한 parameter 인 $n = 1$ 과 $\gamma = 1.0$ 는 이전에 이루어진 실험 구조에 대하여 고리 구조를 예측하기 위해 colony energy 를 적용하는 연구에서 결정된 것을 그대로 이용하였다.[15]

결론 및 토의

전체적 결과는 **Table 1**에 나타내었다. 각 단백질 구조 모델의 고리 부분에 대한 개별적 결과는SI1에서 볼 수 있다. Colony energy, 또는 Galaxy energy가 가장 낮은 구조와 결정된 실험 구조와의 고리 부분의 RMSD 는 colony energy 를 이용했을 때 평균적으로 0.209 Å만큼, 중앙 값으로는 0.815 Å좋아졌다. Sampling 된 구조들을 실험 구조와의 RMSD 값으로 정렬했을 때 에너지가 가장 낮은 구조의 순위는 colony energy 를 적용했을 때 평균적으로 좋아졌다. 고리 부분의 RMSD 와 에너지 값의 Pearson 상관 계수 값은 colony energy 를 적용했을 때 평균적으로 0.006 증가하고 중앙값은0.001 증가했다. S1을 보면 예측된 단백질의 전체 삼차 구조의 고리 구조 예측 후 정확도를 실험 구조와의 전체 RMSD로 비교해보아도 Colony Energy를 적용한 고리 구조로 부분적

Table 1.

	Energy only				**Colony Energy**				
		Lowest energy		min10[5]		Lowest colony energy		min10[6]	RMSD NAT vs TBM[2]
	Corr[3]	RMSD	Rank[4]	best	corr	rmsd	rank	best	
mean	0.057	4.970	211.45	3.8714	0.063	4.761	189.5	4.04	5.374
median	0.016	4.155	158	3.449	0.017	3.634	197	3.573	4.855

7) Score 가 가장 좋은 10개의 구조 중 가장 작은 RMSD

2) 실험 구조와 예측된 전체 단백질 TBM 구조 모델 사이의 RMSD, 즉 구조 모델의 정확도

3) 고리 부분들에 대하여 선택된 고리 부분의 RMSD 와 에너지 값의 상관 계수

4) 선택된 구조의 RMSD 상으로 정렬했을 때 순위

인 정밀화가 이루어졌을 때 평균적으로 0.404Å, 중앙 값은 0.700Å좋아진다. 이와 같은 결과들을 토대로 단백질 구조 모델 정밀화를 위한 고리 구조 예측을 수행할 때 자유에너지를 고려하도록 colony energy 개념을 도입한 것이 예측 구조의 정밀화를 도왔다고 볼 수 있다.

하지만 기존에 실험 구조에 대해 이루어진 colony energy 적용 연구 결과[15]와 비교하며 절대적인 관점에서 결과를 볼 때 자유에너지를 고려하는 것이 단백질 고리 구조 예측 측면에서 많은 향상을 주지 못하고 한계를 보이는 이유는 몇 가지로 예상해볼 수 있다. 먼저, colony energy 를 계산하는 과정만 생각해볼 때 계산을 위한 parameter 를 기존 실험 구조에 대해 training 된 것을 그대로 사용한 과정에 문제가 있을 수 있다. Colony energy 계산 방법 자체를 예측해야 하는 고리 부분의 환경이 부정확해 훨씬 더 어려운 문제를 제공하는 단백질 구조 모델을 대상으로 최적화해야 될 필요가 있다고 생각된다. 한편으로, colony energy 를 계산할 때에도 물리화학 기반의 Galaxy energy 값이 반영되는데 예측된 구조 모델과 같은 부정확한 환경에서 유발되는 오류로 인해 고리 부분에 대한 에너지 계산을 통한 scoring 도 어려울 것으로 생각된다. 따라서 곁가지 최적화뿐 아니라 주변 환경의 backbone까지 최적화해야 할 필요성도 있을 것이다. 또한, 자유 에너지를 고려하는 것은 scoring 단계에서만 적용되는 것이기 때문에 예측된 단백질 구조 모델 환경에서 정확한 고리 구조 sampling 이 충분히 되지 못했다면 scoring 결과에도 영향을 미치게 될 것이다.

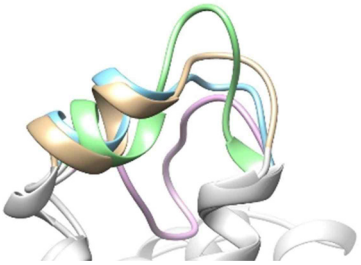

Figure 10. Colony energy 를 적용한 성공적인 사례의 하나. 실험 구조는 카키색, 가장 낮은 colony energy 를 갖는 구조는 하늘색, Galaxy energy 가 가장 낮은 구조는 초록색, 정밀화하기 전 예측된 TBM 구조는 분홍색이다.

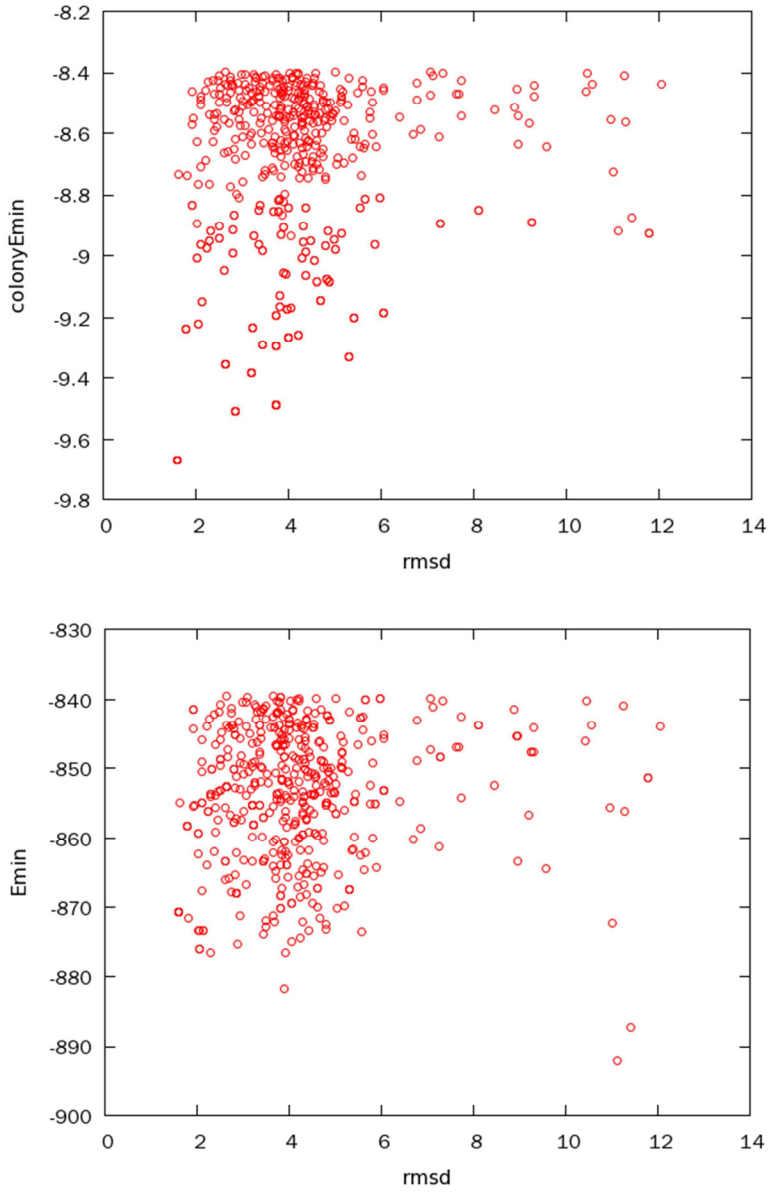

Figure 2. Score-RMSD correlations, 위는Colony Energy, 아래는 Only Energy

가장 성공적인 예시의 하나로 PDB ID 가 4wji 인 단백질에 대한 구조 예측 및 정밀화 결과를 **Figure 1**. 을 통해 보였다. Colony energy 가 가장 낮은 구조가 실험 구조와의 RMSD도 sampling 된 구조 중 가장 좋다. TBM 예측 구조에서 RMSD 가 4.161 Å 이었던 고리 부분은 고리 구조 예측 후 colony energy 가 가장 낮은 구조의 RMSD 가 1.603 Å 로 2.558 Å이 좋아지며 구조 모델을 정밀화할 수 있었다. 반면, Galaxy energy 를 이용해 scoring 했을 때에는 고리 구조의 정확도가 더 낮았다. 예측된 고리 구조들의 정확도와 에너지 값의 Pearson 상관 계수 역시Colony energy 를 적용했을 때 0.005에서0.026으로 좋아졌다. **Figure 2**. 에서 각 에너지 계산 방법과 고리 구조의 정확도의 상관관계를 표현하였다.

결론

이번 연구에서는 기존 연구에서 실험 구조에 대한 고리 구조 예측을 할 때 자유 에너지를 고려하기 위해 colony energy 개념을 도입한 것을 연장시켜 주형 구조를 기반으로 예측된 구조 모델을 정밀화하기 위한 고리 구조 예측에colony energy 를 이용하였다. 주형 기반 구조 모델을 정밀화하기 위한 고리 구조 예측은 고리 부분의 환경의 부정확도로 인해 기존에 scoring 을 위해 많이 사용되는 물리화학 에너지 함수를 기반으로 위치 에너지만을 계산할 때 많은 오류를 포함하게 된다. 이러한 문제로 고리 구조의 Sampling 뿐만 아니라 scoring 에서도 한계가 있게 된다. Colony energy 개념을 도입하여 Sampling 된 고리 구조들의 엔트로피까지 고려하게 되면 이러한 어려운 난이도의 문제에서 어떤 효과를 보이는 지 알아보기 위하여 이번 연구를 수행하였다. 큰 향상을 보이는 몇 가지 사례가 있었지만 일반적으로 정확도를 많이 높이기에는 여전히 한계가 있다는 것을 알게 되었다. 하지만 고리 구조의 sampling 단계에서부터 주형 단백질 구조 모델의 부정확한 환경을 고려하고 colony energy 계산 방법도 최적화 시키는 차후 연구가 이루어진다면 Colony energy 를 적용하는 단백질 구조 모델 정밀화에 더 발전이 있을 것으로 기대가 된다.

감사의 글

본 논문은 2015년도 정부(미래창조과학부)의 재원으로 한국연구재단 첨단 사이언스•교육 허브 개발 사업의 지원을 받아 수행된 연구(No. NRF-2012-M3C1A6035357)로부터 작성되었다.

참고문헌

1. Xiang, Z.; Soto, C. S.; Honig, B. *Proc. Natl. Acad. Sci. U. S. A.***2002**, *99*, 7432.
2. Taylor, J. C.; Takusagawa, F.; Markham, G. D. *Biochemistry***2002**, *41*, 9358.
3. Murre, C.; McCaw, P. S.; Vaessin, H.; Caudy, M.; Jan, L. Y.; Jan, Y. N.; Cabrera, C. V.; Buskin, J. N.; Hauschka, S. D.; Lassar, A. B.; Weintraub, H.; Baltimore, D. *Cell***1989**, *58*, 537.
4. Auffinger, P.; Bielecki, L.; Westhof, E. *Chem. Biol.***2003**, *10*, 551.
5. Lee, G.-R.; Shin, W.-H.; Park, H.-B.; Shin, S.-M.; Seok, C.-O. *Bull. Korean Chem. Soc.***2012**, *33*, 770.
6. Spassov, V. Z.; Flook, P. K.; Yan, L. *Protein Eng. Des. Sel.***2008**, *21*, 91.
7. Soto, C. S.; Fasnacht, M.; Zhu, J.; Forrest, L.; Honig, B. *Proteins***2008**, *70*, 834.
8. Fiser, A.; Sali, A. *Bioinformatics***2003**, *19*, 2500.
9. Lee, J.; Lee, D.; Park, H.; Coutsias, E. A.; Seok, C. *Proteins***2010**, *78*, 3428.
10. Canutescu, A. A.; Shelenkov, A. A.; Dunbrack, R. L. *Protein Sci.***2003**, *12*, 2001.
11. Heo, L.; Park, H.; Seok, C. *Nucleic Acids Res.***2013**, *41*, W384.
12. Shin, W.; Lee, G.; Heo, L.; Lee, H.; Seok, C. *Bio Design*, **2014**, 2 (1), 1-11
13. Park, H.; Seok, C. *Proteins***2012**, *80*, 1974.
14. Lee, J.; Seok, C. *Proteins***2008**, *70*, 1074.
15. Kang, B.;Lee, G.; Seok, C, *제4회 Edison 논문경진대회*, **2014**

S1 각 고리 부분별 자료

Target	Energy only			Colony Energy			RMSD	Env
	Corr	RMSD	Rank	Corr	RMSD	Rank	NAT vs TBM	RMSD
T0762-D1_1	0.009	5.262	158	0.008	3.634	20	7.594	1.841
T0764-D1_1	0.067	13.014	150	0.087	13.014	150	10.768	2.672
T0764-D1_4	0.002	6.522	379	0.012	5.715	199	5.15	2.616
T0776-D1_2	0.002	10.446	130	0.009	11.651	282	10.612	2.662
T0801-D1_1	0.025	2.702	266	0.001	2.519	141	2.586	1.197
T0801-D1_3	0.249	2.029	399	0.512	1.835	209	3.655	0.955
T0801-D1_4	0.032	1.578	250	0.078	1.384	112	1.717	1.849
T0807-D1_2	0.025	5.93	88	0	7.846	220	7.604	2.531
T0811-D1_2	0.003	2.885	128	0.009	3.113	295	3.142	1.532
T0813-D1_2	0.005	11.11	494	0.026	1.603	1	4.161	2.047
T0815-D1_2	0.007	2.559	137	0.013	2.672	213	3.928	1.306
T0815-D1_3	0.07	1.985	17	0.007	5.488	461	4.09	2.028
T0817-D2_1	0.326	2.702	89	0.299	2.702	89	2.459	1.977
T0817-D2_2	0.192	3.531	149	0.005	3.527	148	4.142	1.287
T0819-D1_3	0.002	3.704	342	0.026	2.325	2	5.442	2.832
T0840-D1_1	0	8.49	285	0	10.009	387	11.126	1.146
T0840-D1_2	0	3.605	172	0.006	3.314	63	4.855	0.574
T0843-D1_1	0.175	1.23	140	0.171	1.211	121	3.527	2.406
T0843-D1_2	0.001	7.3	370	0.006	3.928	34	7.973	0.789
T0843-D1_3	0.104	6.016	50	0.107	6.016	50	5.16	1.149
T0851-D1_1	0.015	2.762	213	0.067	2.807	251	6.227	1.718
T0851-D1_2	0.023	5.098	336	0	5.101	338	5.812	1.85
T0851-D1_3	0.019	7.327	378	0.017	6.087	67	7.144	1.123
T0851-D1_4	0	2.93	322	0.095	2.799	197	3.56	1.584
T0854-D2_1	0.002	5.806	480	0.02	5.806	480	5.672	2.167
T0856-D1_2	0.052	4.548	17	0.067	6.089	338	7.136	2.442
T0856-D1_3	0.102	3.882	76	0.039	3.573	39	4.529	2.994

T0856-D1_5	0.012	4.155	118	0.005	4.658	216	3.833	1.698
T0858-D1_1	0.016	4.65	4	0.014	6.603	254	4.351	1.557
T0858-D1_3	0.216	3.633	115	0.215	3.633	115	2.929	2.52
T0858-D1_4	0.002	6.694	303	0.042	6.918	383	5.696	2.263
mean	0.05661	4.97048	211.45	0.0633	4.7606	190	5.373548387	1.848774
median	0.016	4.155	158	0.017	3.634	197	4.855	1.849