2022년 3월 8일 CompBio 회의록

- 구글 시트로 그동안 논의되었던 논문 정리해서 업로드

- 병혁)데이터 백업 -> 플랜, 컴퓨팅 리소스 관리 -> 각자 나눠서 관리(관리자: 박사과정)

- 주간레포트는 병혁ppt, 나머지는 깃허브 markdown에 한 topic 마다 쭉 연결해서 이어놓을 것. source code나 관련 자료 모두 하나의 topic에 해당하는 repository (code는 google colab이용)

- 매주 이렇게 미팅하는 것으로 진행할 것.

- 가지고 있는 topic, idea, 진행 상황 보고

<병혁>

topic:

* PUL regulation paper(duleepba와 논의중),

* PULFinder: 현주와 HMM model로 하는 것, R로는 이전에 구현했는데 python에서 pfam, kofam, -> ClusterFinder로 쓰면 안되나? code를 그대로 쓰면, input만 Clusterfinder가 읽을 수 있게 하는 quick test, clusterfinder backbond에 input만 바꿔서 넣을 수 있게 바꾸는 법.

=> quick test를 통해서 data가 원하는 대로 나오는지 보는거라(R로는 했음)

-> HMM package가 update가 안되거나, fitting해서 쓰는 게 위주여서 아예 새로 코드를 짰음

=> time line, model training하고 clean data 관리(~3월 말) publication까지 목표

probiotics과제 관련해서 gene cluster 분석 파트를 share 할 예정

정준 – genome refseq NCBI에서 가져오는 것 해서 균주 400개 정도 다운 받아놨음.

각 속별로 어떤 기능성 관련 gene cluster가 분포되어 있는지, 특정 속에는 어떤 특징을 가지고 있는지 해석하는 것. 일단은 pfam으로 training 시켜서 할 수 있는 것.

deepBGC는 기존에 알려져 있는 것에서 찾는 것(knowledge-based) 유산균에서 관심있어 하는 것은 세균 밖에서의 생합성에 대한 것. sanger-polysaccharide synthesis 같은 것들. PUL은 분해관련도 있지만 합성에만 집중해서 accessary 같은 것들 KEGG에서 metabolic pathway 들도 있으니까,

=> 1. 먼저, deepBGC에 갖고 싶은 data 관련해서 있으면 그대로 training data로 prediction하면 되고,

=> 2. 알려진 게 없다 하면 FungCluster 하듯이 evolutionary conserved gene cluster로 해서 gene cluster prediction한 다음에 그중에 polysaccharide 관련 gene cluster를 뽑아내는 것.

PULFinder Pfam + KG로 KG로 우산을 씌우고 그 밑에 Pfam이 들어가는 형태로 하면 더 정확하게 나올 것. DBcan?에서 400개 정도 데려와서 gene cluster 분석을 하면 될 것

=> deepBGC는 기존 분석 밖에 없으니까 이건 impact가 없음. 이걸 meaningful하게 따로 추출을 하는 형태로 변환해야함.

=> target하는 것은 toxic gene이 있는지, antibacterial 등등 사람들이 궁금해 하는 정보를 모두 담고 있는 database로 이용

결국은 probiotics나 gut microbiome 등등 biosynthetic polysaccharide에 집중하는 이유는 세균 밖에 있는 polysaccharide를 redesign해서 해당 probiotics와 gut microbiome의 기능과 연관지어서 cell-surface polysaccharide로 typing할 수 있으면 functional classification을 해보자! (immune의 강/중/약으로 분류)

그걸 하기 위해서 먼저 deepBGC같은 tool을 사용해보자 하는 것이니까.

[exploratory analysis] + 잡학지식 습득, meta data까지 handling

대사체 과제는 RNAseq (업체 선정 후 contact, sample data 바로 보낼 수 있게 연락.) -> handling

카드는 activation 했음 -> 알아서 deposit 하기.

- 감귤 동충화초는 다음주 16일에 온라인 회의, 책임자만 들어가니까 신경x

결과는 1~2주 이내로 나올 것이고, microbiome study 용.

- 기초연구실 3/10, 3/15 쯤에 온라인 회의.

<보근>

- FungCluster: 어제 발표에 더해서

Evolcluster 논문 2개 -> Nature를 모델로 삼고, 어떤 것을 어떻게 분석해서 어떤 결과가 나온 것인지를 정리하고, fungi도 deepBGC 돌릴 수 있다고 함. -> benchmark 상에서의 비교를 확실하게. FungCluster가 어떤 장점을 가지고 있는지 강조, flow에 따라서 result와 evaluation까지 진행. (4만 개 중에 제대로 나온 것이 몇 개인지 확인하고, sorting해서 평가표를 만들 것.) -> 기준을 정해서 어떤 것이 좋고 나쁜지에 대한 평가가 가능할 것.

Members 수가 일정하지 않음 -> 합치는 과정 상의 문제 -> filter를 하려면? -> 직접 merge하고 분리하면서 확인해볼 것, 이 과정은 benchmark가 필요 없음(Final이랑 비교하는 용도).

data돌리고, clean up해서 result 완성하고 benchmark할 것. -> case study

Manually finding와는 다른 특징적인 부분을 발견해서 refind할 것.

여기에 machine learning algorithm을 넣을 수도 있는지?

병혁) 4만 개를 다 뽑아서 어떤 parameter가 어떤 특성들을 가지고 있는지를 확인해봐야 하는데, number of genome, number of gap이 중요 -> statistics로 분류했을 때 다시 merge하면서 score를 세울 수 있을 것. (length blast 같은 것과 동일) -> quality check, 3D로 gap number까지 확인하던가(특정 부분을 section화 시키기)

병혁) 9만개의 genes에서 4만개의 cluster로 assign 된 게 있으면, 각각의 gene이 gene cluster 상에서 얼마나 중복되는지, 어느 그룹에 속하는지를 확인하면, 어떤 gene을 중요하게 생각하고, 어떤 gene을 무시해도 되는지를 결정할 수 있을 것. -> 잘 쓰이는 gene 들의 ortholog 로 되어 있는 cluster를 중요하게 생각하되 중복을 없애는 대상이기도 함. -> gene cluster로 분석된 것들의 onthology를 다시 분석했을 때 cluster의 의미를 다시 확인할 수 있을 것(global analysis).

scaffold가 많이 나뉘어져 있어서 order를 고려하지 않는다는 장점이자 단점임. 순서 상관없이 다 잡아낼 수 있지만 alignment 관련해서는 단점으로 작용할 수도 있음.

9개 genome에서 가장 conserved 되어 있는 것들만 고려하겠다. 이렇게 data를 좁혀서 해석할 수도 있음.

ortholog로 시작하기 때문에 (Evol: sequence homology, 다른 것들도 homolog) 근데 일단 4만개의 성질을 눈으로 보고 평가를 진행해서 정말 step에 필요한 게 있는 것들만 확인해서 넘어가는 것.

4만 8천개가 정말 reasonable한 숫자인지를 확인해야함.

orthoMCL clustering = network partition MCL하는 것. 그냥 MCL로 partitioning하기. Network의 feature도 봐야함. partitioning의 문제인지, 아니면 다른 것의 문제인지

일본, linkage clustering -> node 수가 많아서 시간이 오래 걸리므로 진행하지 않았었음. data가 크지 않았는데 오래 돌렸음. high로 잡고 threshold로 잘라 냄.

기존 논문의 원리를 이용하면서 속도를 올릴 수 있는 다른 algorithm이 있으면 사용할 수도 있을 것 -> cluster하고 cut하는 과정

- intergenic ORF finder 마무리

fungi의 intergenic length를 볼 것. length가 길다 = annotation missing, transposon

highly conserved 되어 있는 것은 그대로, Lincluster 같은 것으로 압축하고, ORF인지 아닌지를 prediction할 수 있는 machine learning 이용하거나 해서 short repeat이나 conserved part가 있는지를 확인하는 것.

Top10은 이미 known이기 때문에 이게 아니라 highly conserved되어 있는데 등장하지 못한 것들 것 왜 발견되지 못했는지에 대한 이유를 확인하고, alphafold로 structure보고 이걸로 homology 검색하면 function 빠르게 확인할 수 있을 것.

독버섯 -> Mycocosm DB

독버섯 list, function search, taxa에서 annotation -> 독버섯이 아니었는데 독버섯으로 분류된 것들도 존재

toxic mushroom gene DB: antibacterial, toxin, adhesin protein 등의 유무로 safety evaluation tool로 사용할 수 있을 것.

antiSMASH, deepBGC, --- 다 써서 data 취합해서 새로 하나 만들 것.

---

alpha fold PUL -> extracellular 물질에 대한 complex prediction => extracellular만 extract

sequence alignment가 오래 걸려서 rosetta로 scan하고 나서 alpha fold에 이용한 전례가 있음.

1500개 이하는 google colab에서 alpha fold 진행도 해볼 것.