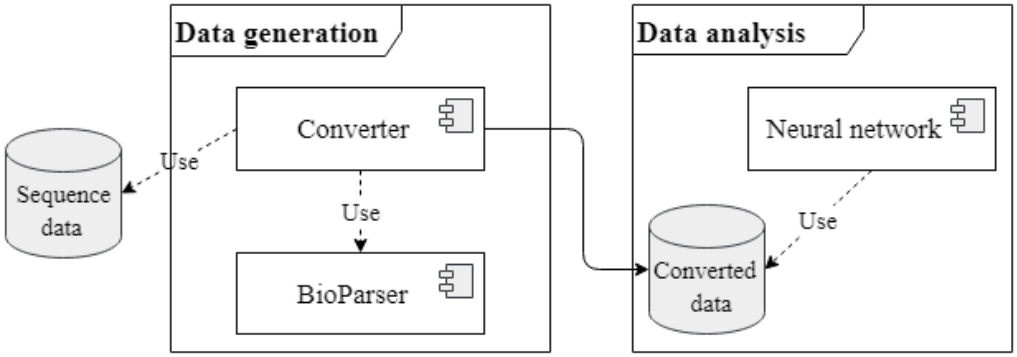
Что есть сейчас (распознавание):

1. Архитектура решения:



2. Sequence data из бд SIIVA (подпоследовательности длины 512)
3. Грамматика:

```
[<Start>]
s1: stem<s0> any
a_0_7 : any*[2..10]
s0: a_0_7 | a_0_7 stem<s0> s0
any: A | U | C | G
stem1<s>:
     A s U
    | G s C
    | U s A
    | C s G
stem2<s>: stem1<stem1<s>>
stem<s>:
      A stem<s> U
    | U stem<s> A
    | C stem<s> G
    | G stem<s> C
    | stem1<stem2<s>>
```

4. Нейронная сеть в файле Model1

	classified as positive	classified as negative
positive	TP = 2789	FP = 856
negative	FN = 108	TN = 3592

- Accuracy $= \frac{TP+NT}{TP+TN+FP+FN} = 0.87$
- Precision $= \frac{TP}{TP+FP} = 0.96$
- Recall $= \frac{TP}{TP+FN} = 0.77$
- Specificity $= \frac{TN}{TN+FP} = 0.97$

Планы (классификация):

1. Последовательности из GreenGenes, только бактерии с полной классификацией
2. Добавить в метаданные информацию из дерева жизни
3. Взять только те организмы, для которых в базе есть 100 и более цепочек. Всего 143 вида, достаточно равномерно распределенных по общему дереву жизни
4. Длина подцепочки?
5. Сбалансированная выборка: чем меньше цепочек есть в базе, тем меньше шаг
6. Улучшить парсер и грамматику
7. Скрипт для подбора параметров модели нейросети и обучения
8. На выходе -- вероятностная модель

Жалкая попытка архитектуры решения: