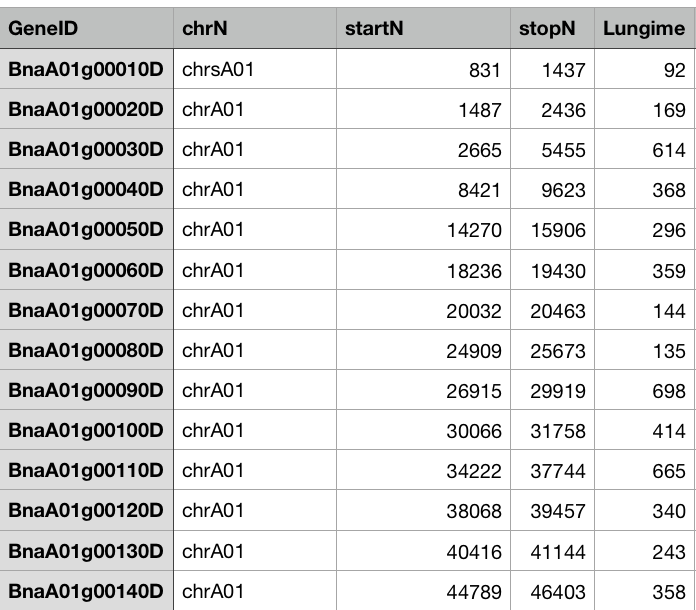
Task 1.

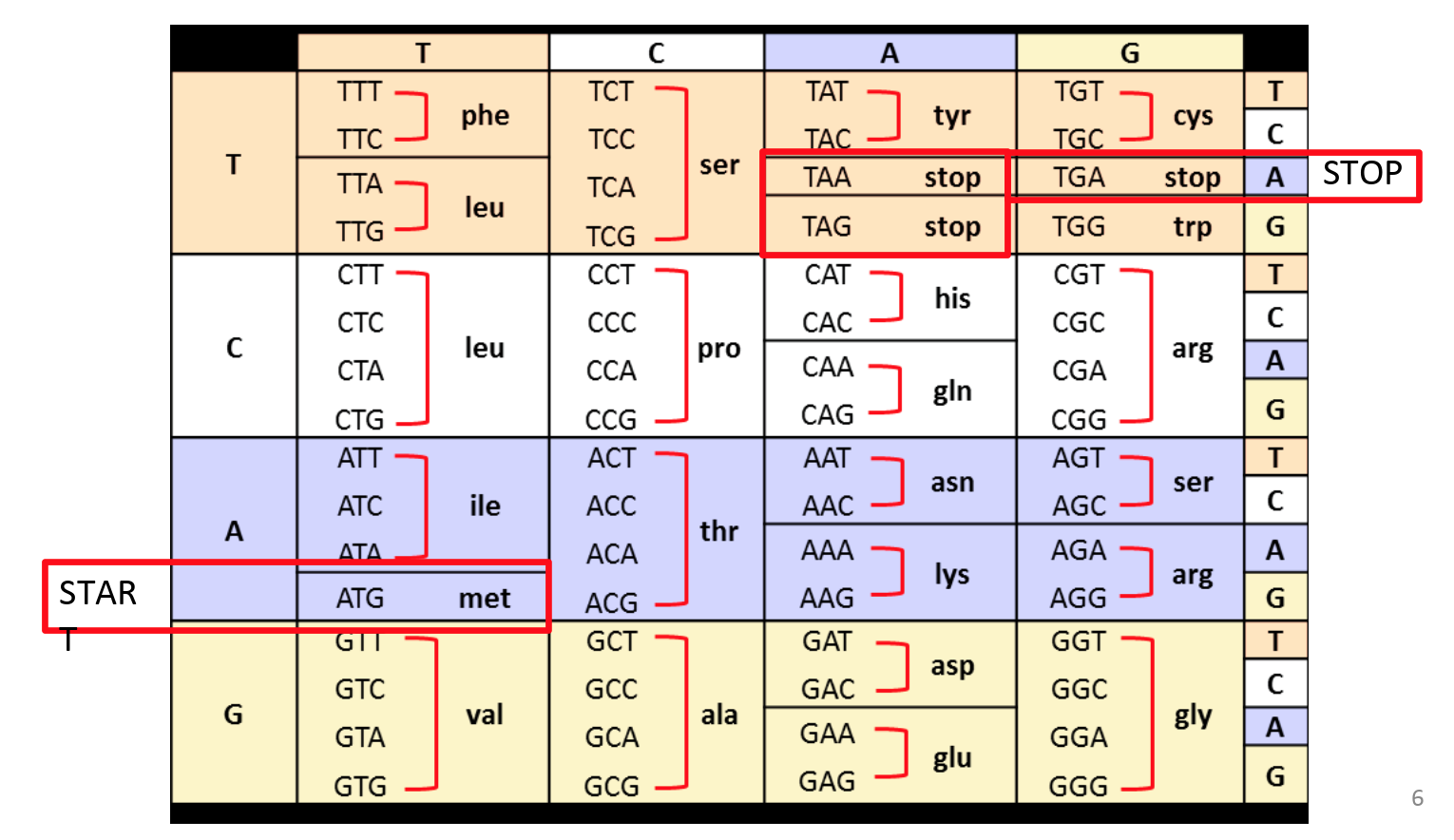
Task 1.1.

Formatul inputului nostru (outputul celor de la segmentare) este urmatorul:

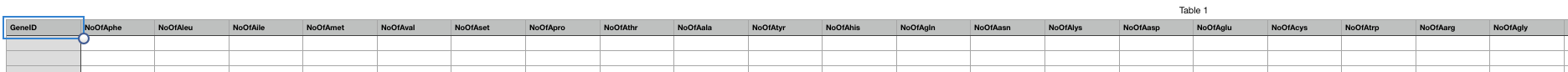


Primul task este sa ne folosim de aceste date pentru a genera un mod optim de reprezentate pentru clasificarea acestora.

Primul pas e sa gasim toti aminoacizii unei anumite gene. Mai jos sunt aminoacizii posibili(20 aminoacizi).



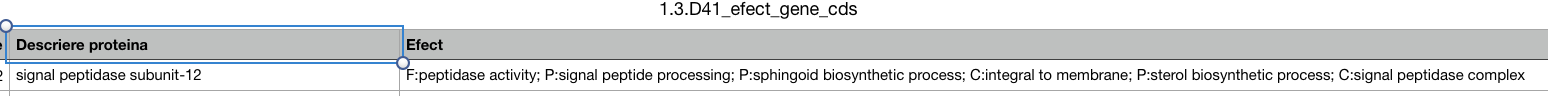
Se va genera un table sub forma:



Pentru fiecare gena se va indica cantitatea fiecarui aminoacid prezent in acea gena.

Task 1.2

Dupa finalizarea Task-ului 1.1, va trebui sa adaugam coloana “Efectul proteinei”(si “Descrierea proteinei”, dar pentru asta nu am gasit datele inca).

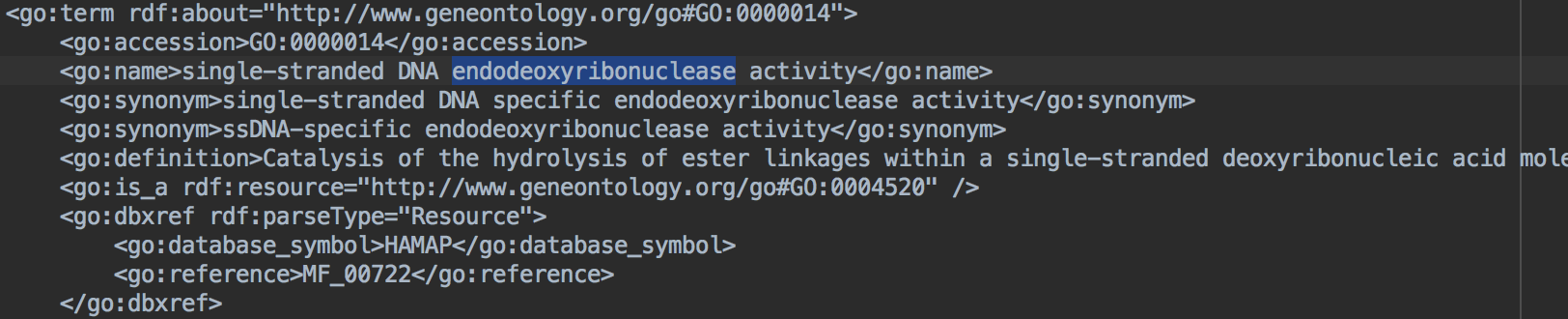


Doar ca ar trebui sa le afisam in formatul codificat de pe geneontology.

De ex. Avem efectul urmator din datele detaliate:



Si avem fisierul in format rdf care contine toate informatiile.



Ceea ce o sa adaugam noi pentru fiecare gena in coloana “Efectul proteinei” va fi tagul <go:accession> : GO:0000014…si restul datelor sub forma GO:id ce sunt gasite in tabel.

Task 2.

Inputul pentru clasificare va fi urmatorul:   
Training Dataset:

GeneId | NoOfamino1 | NoOfamino2| ….. | NoOfamino20 | Effect

BNABLA | 12 | 21 |… | 0 | GO:00012, GO:00014

…

Test Dataset:

GeneId | NoOfamino1 | NoOfamino2| ….. | NoOfamino20

BNABLA | 12 | 21 |… | 0

…

Vom folosi o librarie din python cu care sa antrenam un model. Cred ca putem face 2 metode, respectiv:

Task 2.1

KNN algorithm pentru clasificare

https://scikit-learn.org/stable/modules/neighbors.html

Task 2.2

Decision Tree/ID3 pentru clasificare

https://scikit-learn.org/stable/modules/tree.html