Προγραμματισμός με τη γλώσσα python

Alexandros Kanterakis kantale@ics.forth.gr (mailto:kantale@ics.forth.gr)

Διάλεξη 12η, Παρασκευή 26 Ιανουαρίου 2018

Σκοπός αυτής της διάλεξης είναι η εισαγωγή σε μεθόδους για να προσπελαύνουμε υπάρχοντες online βάσεις δεδομένων οι οποίες περιέχουν γενετικές-βιολογικές πληροφορίες.

Καταρχήν πρέπει να αναφέρουμε ότι υπάρχουν χιλιάδες τέτοιες βάσεις οι οποίες περιέχουν πληροφορίες για όλες τις μορφές των -omics data. Μία πολύ καλή λίστα βρίσκεται εδώ: https://www.oxfordjournals.org/nar/database/a/ (https://www.oxfordjournals.org/nar/database/a/) (προφανώς και η λίστα δεν είναι πλήρης).

Πως λοιπόν μπορούμε προγραμματιστικά να προσπελάσουμε αυτές τις βάσεις;

HTTP GET and POST requests

Πριν απαντήσουμε σε αυτό πρέπει να πούμε κάποια βασικά στοιχεία για το HTTP (https://en.wikipedia.org/wiki/Hypertext_Transfer_Protocol) πρωτόκολλο. Το HTTP είναι ένα πρωτόκολλο που καθορίζει πως δύο υπολογιστές μπορούν να επικοινωνήσουν ώστε ο ένας να "πάρει" κάποια πληροφορία από τον άλλο. Αυτή η "πληροφορία" είναι συνήθως μία HTML σελίδα, αλλά μπορεί να είναι και οποιαδίποτε άλλη μορφή πληροφορίας. Μία πηγή πληροφορίας ορίζεται σύμφωνα με το HTTP ως ένα URL ή αλλιώς ένα link. π.χ: http://bioinfo-grad.gr/eclass/ (http://bioinfo-grad.gr/eclass/).

Ένας υπολογιστής που "ζητάει" πληροφορία συνήθως αναφέρεται ως "client" ενώ ένας υπολογιστής που μπορεί να τη δώσει ονομάζεται ως server. Ένας client για να "ζητήσει" πληροφορία από τον "server" πρέπει να ξέρει το URL του server. Στη συνέχεια εκτελεί ένα "request" (υποθέτω "αίτημα" στα ελληνικά), προς τον server και του δηλώνει ποια είναι η μορφή της πληροφορίας που θέλει. Ο server αποφασίζει αν την έχει (και αν πρέπει να τη δώσει) την πληροφορία και ανάλογα την επιστρέφει στον client (κάτι σαν το ΚΕΠ..). Σε περίπτωση που κάτι δεν πάει καλά, ο server μπορεί να ενημερώσει τον client τι λάθος έγινε επιστρέφοντας έναν κωδικό (https://en.wikipedia.org/wiki/List_of_HTTP_status_codes). Αν ο κωδικός είναι το 200, τότε δεν συνέβει κανένα λάθος. Όταν ένας server "απαντάει" σε ένα request, τότε λέμε ότι έστειλε ένα responce.

Όταν ο client επικοινωνεί με τον server του δηλώνει: Τι θέλει, και σε ποια μορφή το θέλει.

Σχετικά με το "τι θέλω" υπάρχουν ΔΥΟ βασικοί τρόποι δήλωσης: οι "GET" και "POST" (http://www.w3schools.com/tags/ref_httpmethods.asp).

Η GET είναι ένας τρόπος να δηλώσεις τι θέλεις από έναν server, κωδικοποιώντας το αίτημα σου πάνω στο link. Αν υποθέσουμε π.χ. ότι θέλεις πληροροφίες για το προιόν με id=123 και το οποίο ανοίκει στη κατηγορία category=678 τότε το link που εκτελεί ένα GET request θα έχει τη μορφή: http://www.example.com/index.html?id=123&category=678 (http://www.example.com/index.html?id=123&category=678). Για παράδειγμα αν κοιτάξουμε αυτό το εντελώς τυχαίο link: http://news.in.gr/greece/article/?aid=1500029633 (http://news.in.gr/greece/article/?aid=1500029633) θα δούμε ότι είναι ένα GET request στη σελίδα news.in.gr/greece/article όπου ζητάει το "resource" το οποίο έχει aid=1500029633. Το GET request είναι ο κύριος τρόπος επικοινωνίας των browsers με τους web servers.

Τα GET requests έχουν δύο μειονεκτήματα:

- Αυτό που ζητάει ο client "φαίνεται" πάνω στο link. Γενικότερα τα links δεν θεωρούνται "ασφαλή" πληροφορία και είναι προσβάσιμα από πολλούς "ενδιάμεσους". Αν για παράδειγμα ζητάμε κάποια δεδομένα με βάση προσωπικές μας πληροφορίες (π.χ. τηλέφωνο) δεν είναι καλή ιδέα να φαίνεται αυτό στο link.
- Δεδομένου ότι αυτό που ζητάμε βρίσκεται πάνω στο link. Αν αυτό που ζητάμε είναι πολύπλοκο, τότε το link μπορεί να γίνει τεράστιο και μη διαχειρίσιμο. Για παράδειγμα να ένα άσχημο link με πολλά

GET πεδία: https://www.amazon.com/Data-Visualization-Python-JavaScript-Transform/dp/1491920513/ref=s9_cartx_gw_g14_i5_r?_encoding=UTF8&fpl=fresh&pf_rd_m=ATVPDKIKX0DER&pf_rd_s=&pf_rd_r=WRMBPSF9K3H62TEWG5DW&pf_rd_t=36701&pf_rd_p=a6aaf593-1ba4-4f4e-bdcc-0febe090b8ed&pf_rd_i=desktop (https://www.amazon.com/Data-Visualization-Python-JavaScript-Transform/dp/1491920513/ref=s9_cartx_gw_g14_i5_r?_encoding=UTF8&fpl=fresh&pf_rd_m=ATVPDKIKX0DER&pf_rd_s=&pf_rd_r=WRMBPSF9K3H62TEWG5DW&pf_rd_t=36701&pf_rd_p=a6aaf593-1ba4-4f4e-bdcc-0febe090b8ed&pf_rd_i=desktop)

Ο δεύτερος τρόπος επικοινωνίας του client με τον server είναι μεσώ του POST request. Όταν κάνουμε POST βάζουμε μέσα σε ένα ειδικό πεδίο τα δεδομένα μας και το HTTP πρωτόκολο στέλνει αυτά τα δεδομένα στον server χωρίς να φαίνονται στο link. Π.χ. όταν κάνετε login στο site: http://bioinfo-grad.gr/eclass/ (http://bioinfo-grad.gr/eclass/) βάζετε το username και το password σας. Στη συνέχεια ο browser στέλνει αυτά τα δεδομένα στον server για επιβεβαίωση. Αυτά τα δεδομένα ΔΕΝ πρέπει προφανώς να φαίνονται στο link. Πως γίνεται αυτό; Αν κοιτάξετε τον κώδικα της σελίδας (π.χ. από Firefox μπορείτε να το δείτε επιλέγοντας Tools -> Web Developer -> Page Source). Θα δείτε σε ένα σημείο να γράφει:

```
<h2>Σύνδεση χρήστη</h2>
<div><form action='http://bioinfo-grad.gr/eclass/' method='post'>
```

Δηλαδή λέει: τα δεδομένα που βάζουμε σε αυτή τη φόρμα (form) θέλω να στέλνονται στον server μέσω POST request.

Ένα τελευταίο σημείο για το HTTP είναι οι headers (https://en.wikipedia.org/wiki/List_of_HTTP_header_fields). Τα headers (επικεφαλίδες;;) είναι πεδία με προ-καθορισμένα ονόματα στα οποία μπορούμε να βάλουμε πληροφορίες που διευκολύνουν τον server να διεκπεραιώσει το request μας. Π.χ. αν βάλουμε στο header με το όνομα "Content-Type" τη τιμή "application/json", τότε λέμε στον server ότι τα δεδομένα που του στέλνουμε μέσω POST είναι σε μορφή JSON.

python requests package

Ωραία όλα αυτά, αλλά πως τα κάνουμε μέσω python; Παρόλο που η python έχει βιβλιοθήκες για να κάνουμε GET και POST, θα χρησιμοποιήσουμε ένα εξωτερικό πακέτο το οποίο είναι εξαιρετικά εύχρηστο. Αυτό το πακέτο είναι το requests: http://docs.python-requests.org/en/master/ (http://docs.python-requests.org/en/master/) . Για να το εγκαταστίσετε τρέξτε:

```
pip install requests
```

ΠΡΟΣΟΧΗ! πρέπει να βεβαιωθείτε ότι το πρόγραμμα pip βρίσκεται στο ίδιο σημείο που είναι και η python που τρέχετε!. Για παράδειγμα:

```
$ which pip
/Users/alexandroskanterakis/anaconda3/bin/pip
$ which python
/Users/alexandroskanterakis/anaconda3/bin/python
```

Βλέπουμε δηλαδή ότι το pip και η python που τρέχω είναι στο ίδιο σημείο. Διαφορετικά θα πρέπει να γράψετε κάτι σαν αυτό:

```
/Directory/where/your/python/is/pip install requests
```

Για να βεβαιωθείτε ότι έχει εγκατασταθεί σωστά θα πρέπει να μπορείτε να κάνετε import αυτό το πακέτο χωρίς πρόβλημα:

```
$ python
Python 3.5.2 |Anaconda 4.1.1 (x86_64)| (default, Jul  2 2016, 17:52:12)
[GCC 4.2.1 Compatible Apple LLVM 4.2 (clang-425.0.28)] on darwin
Type "help", "copyright", "credits" or "license" for more information.
>>> import requests
>>>
```

Προσπελαύνοντας την ENSEMBL

Η ENSEMBL (http://www.ensembl.org/index.html) είναι ένα από τα πιο σημαντικά portals με διάφορες γενομικές πληροφορίες για πολλούς οργανισμούς. Η ENSEMBL έχει ένα API (Application programming interface (https://en.wikipedia.org/wiki/Application_programming_interface)) το οποίο είναι ένα σύνολο από οδηγίες προς προγραμματιστές για να προσπελαύνουν μία υπηρεσία. Το API της ENSEMBL περιγράφεται εδώ: http://rest.ensembl.org/ (http://rest.ensembl.org/)

Παρατηρήστε ότι κάποιες υπηρεσίες είναι προσβάσιμες μέσω GET και κάποιες μέσω POST.

Σαν παράδειγμα θα χρησιμοποιήσουμε το API της ENSEMBL το οποίο αναφέρεται στο Variant Effect Predictor (http://www.ensembl.org/info/docs/tools/vep/index.html). Το Variant Effect Predictor μας δίνει διάφορες πληροφορίες σχετικά με το downstream effect και clinical significance ενώς variant. Αλλα πριν κάνουμε αυτό ας δούμε με ποιους τρόπους κωδικοποιούμε έναν variant.

Γενικότερα υπάρχουν δύο τρόποι.

Αν ο variant είναι γνωστός τότε κοιτάμε αν έχει καταχωρηθεί στη dbSNP (https://www.ncbi.nlm.nih.gov/projects/SNP/). Σε αυτή τη περίπτωση ο variant θα έχει εναν κωδικό με το φορμάτ: rsXXXXXX. πχ: rs56116432.

Αν ο variant δεν είναι γνωστός και δεν υπάρχει στη dbSNP τότε μπορούμε να τον περιγράψουμε μέσω του HGVS (http://varnomen.hgvs.org/recommendations/general/) φορμάτ. Η περιγραφή αυτή είναι αρκετά πολύπλοκη και ξεφεύγει λίγο από τους σκοπούς αυτής της διάλεξης. Απλά θα πούμε ότι μία από τις πολλές HGVS μορφές ενώς variant είναι το:

{ΧΡΩΜΟΣΩΜΑ}:g.{ΘΕΣΗ}{REFERENCE}>{ALTERNATIVE}

π.χ.:

9:g.22125504G>C

Δηλαδή στο χρωμόσωμα 9 στη θέση 22125504 αντί για G που είναι το reference υπάρχει C.

GET request για dbSNP variant

Για να πάρουμε πληροφορίες από το Variant Effect Predictor για ένα dbSNP variant, η ENSEMBL δίνει αυτό το API: http://rest.ensembl.org/documentation/info/vep_id_get (http://rest.ensembl.org/documentation/info/vep_id_get) Μπορούμε να κάνουμε ένα request σε αυτό το API ως εξής:

In [1]:
```
import requests

# Το URL που πρέπει να χρησιμοποιήσουμε
# υπάρχει στη σελίδα http://rest.ensembl.org/documentation/info/vep_id_get
# Επίσης Αφού είναι GET Request βάζουμε τη πληροφορία που θέλουμε στο URL:
url = 'http://rest.ensembl.org/vep/human/id/rs56116432?'
headers = { "Content-Type" : "application/json"}

# Κάνουμε το GET request
r = requests.get(url, headers=headers)
```

Στη συνέχεια κοιτάμαι αν όλα πήγαν οκ:

In [33]:
```
print (r.ok) # Στην ουσία εδώ κοιτάει αν το response έχει κωδικό 200
```

```
True
```

Παίρνουμε τα JSON δεδομένα:

In [10]:
```
data = r.json()
print (data)
```


```
[{'transcript_consequences': [{'biotype': 'processed_transcript',
'gene_id': 'ENSG00000175164', 'cdna_end': 700, 'hgnc_id': 'HGNC:79
', 'gene_symbol_source': 'HGNC', 'cdna_start': 700, 'variant_allel
e': 'T', 'transcript_id': 'ENST00000453660', 'consequence_terms':
['non_coding_transcript_exon_variant', 'non_coding_transcript_vari
ant'], 'impact': 'MODIFIER', 'strand': -1, 'gene_symbol': 'ABO'},
{'biotype': 'protein_coding', 'gene_id': 'ENSG00000175164', 'sift_
score': 0, 'hgnc_id': 'HGNC:79', 'polyphen_score': 0.997, 'sift_pr
ediction': 'deleterious', 'variant_allele': 'T', 'cds_start': 686,
'codons': 'gGc/gAc', 'transcript_id': 'ENST00000538324', 'cds_end
': 686, 'polyphen_prediction': 'probably_damaging', 'impact': 'MOD
ERATE', 'cdna_end': 711, 'gene_symbol_source': 'HGNC', 'cdna_start
': 711, 'consequence_terms': ['missense_variant'], 'protein_end':
229, 'protein_start': 229, 'strand': -1, 'amino_acids': 'G/D', 'ge
ne_symbol': 'ABO'}, {'biotype': 'protein_coding', 'gene_id': 'ENSG
00000175164', 'sift_score': 0, 'hgnc_id': 'HGNC:79', 'polyphen_sco
re': 0.997, 'sift_prediction': 'deleterious', 'variant_allele': 'T
', 'cds_start': 686, 'codons': 'gGc/gAc', 'transcript_id': 'ENST00
000611156', 'cds_end': 686, 'polyphen_prediction': 'probably_damag
ing', 'impact': 'MODERATE', 'cdna_end': 711, 'gene_symbol_source':
'HGNC', 'cdna_start': 711, 'consequence_terms': ['missense_variant
'], 'protein_end': 229, 'protein_start': 229, 'strand': -1, 'amino
_acids': 'G/D', 'gene_symbol': 'ABO'}], 'most_severe_consequence':
'missense_variant', 'input': 'rs56116432', 'seq_region_name': '9',
'allele_string': 'C/T', 'end': 133256042, 'assembly_name': 'GRCh38
', 'colocated_variants': [{'exac_amr_maf': 0.004932, 'exac_sas_maf
': 0.001639, 'exac_nfe_maf': 0.005339, 'end': 133256042, 'eas_maf
': 0, 'exac_adj_allele': 'T', 'exac_eas_maf': 0, 'exac_fin_maf':
0.02601, 'exac_amr_allele': 'T', 'ea_maf': 0.003809, 'exac_oth_all
ele': 'T', 'minor_allele': 'T', 'sas_allele': 'T', 'amr_allele': '
T', 'exac_afr_allele': 'T', 'start': 133256042, 'ea_allele': 'T',
'allele_string': 'C/T', 'exac_eas_allele': 'T', 'afr_maf': 0, 'amr
_maf': 0.0014, 'sas_maf': 0.001, 'eur_allele': 'T', 'exac_fin_alle
le': 'T', 'strand': 1, 'afr_allele': 'T', 'exac_maf': 0.003022, 'm
inor_allele_freq': 0.0026, 'seq_region_name': 9, 'exac_nfe_allele
': 'T', 'exac_adj_maf': 0.004439, 'exac_afr_maf': 0.0005079, 'aa_m
af': 0.0007102, 'exac_sas_allele': 'T', 'eur_maf': 0.0109, 'aa_all
ele': 'T', 'eas_allele': 'T', 'id': 'rs56116432', 'exac_allele': '
T', 'exac_oth_maf': 0.004926}], 'strand': 1, 'id': 'rs56116432', '
start': 133256042}, {'transcript_consequences': [{'biotype': 'prot
ein_coding', 'gene_id': 'ENSG00000281879', 'sift_score': 0, 'hgnc_
id': 'HGNC:79', 'polyphen_score': 0.997, 'sift_prediction': 'delet
erious', 'variant_allele': 'T', 'cds_start': 689, 'codons': 'gGc/g
Ac', 'transcript_id': 'ENST00000626615', 'cds_end': 689, 'polyphen
_prediction': 'probably_damaging', 'impact': 'MODERATE', 'cdna_end
': 714, 'gene_symbol_source': 'HGNC', 'cdna_start': 714, 'conseque
nce_terms': ['missense_variant'], 'protein_end': 230, 'protein_sta
rt': 230, 'strand': -1, 'amino_acids': 'G/D', 'gene_symbol': 'ABO
'}], 'most_severe_consequence': 'missense_variant', 'input': 'rs56
116432', 'seq_region_name': 'CHR_HG2030_PATCH', 'allele_string': '
C/T', 'end': 133256189, 'assembly_name': 'GRCh38', 'colocated_vari
ants': [{'allele_string': 'C/T', 'seq_region_name': 'CHR_HG2030_PA
TCH', 'end': 133256189, 'minor_allele': 'T', 'strand': 1, 'id': 'r
s56116432', 'minor_allele_freq': 0.0026, 'start': 133256189}], 'st
rand': 1, 'id': 'rs56116432', 'start': 133256189}]
```

In [11]:
```
print (len(data))
```

```
2
```

επέστρεψε 2 εγγραφές. Ας πάρουμε την πρώτη:

In [12]:
```
print (data[0])
```


```
{'transcript_consequences': [{'biotype': 'processed_transcript', '
gene_id': 'ENSG00000175164', 'cdna_end': 700, 'hgnc_id': 'HGNC:79
', 'gene_symbol_source': 'HGNC', 'cdna_start': 700, 'variant_allel
e': 'T', 'transcript_id': 'ENST00000453660', 'consequence_terms':
['non_coding_transcript_exon_variant', 'non_coding_transcript_vari
ant'], 'impact': 'MODIFIER', 'strand': -1, 'gene_symbol': 'ABO'},
{'biotype': 'protein_coding', 'gene_id': 'ENSG00000175164', 'sift_
score': 0, 'hgnc_id': 'HGNC:79', 'polyphen_score': 0.997, 'sift_pr
ediction': 'deleterious', 'variant_allele': 'T', 'cds_start': 686,
'codons': 'gGc/gAc', 'transcript_id': 'ENST00000538324', 'cds_end
': 686, 'polyphen_prediction': 'probably_damaging', 'impact': 'MOD
ERATE', 'cdna_end': 711, 'gene_symbol_source': 'HGNC', 'cdna_start
': 711, 'consequence_terms': ['missense_variant'], 'protein_end':
229, 'protein_start': 229, 'strand': -1, 'amino_acids': 'G/D', 'ge
ne_symbol': 'ABO'}, {'biotype': 'protein_coding', 'gene_id': 'ENSG
00000175164', 'sift_score': 0, 'hgnc_id': 'HGNC:79', 'polyphen_sco
re': 0.997, 'sift_prediction': 'deleterious', 'variant_allele': 'T
', 'cds_start': 686, 'codons': 'gGc/gAc', 'transcript_id': 'ENST00
000611156', 'cds_end': 686, 'polyphen_prediction': 'probably_damag
ing', 'impact': 'MODERATE', 'cdna_end': 711, 'gene_symbol_source':
'HGNC', 'cdna_start': 711, 'consequence_terms': ['missense_variant
'], 'protein_end': 229, 'protein_start': 229, 'strand': -1, 'amino
_acids': 'G/D', 'gene_symbol': 'ABO'}], 'most_severe_consequence':
'missense_variant', 'input': 'rs56116432', 'seq_region_name': '9',
'allele_string': 'C/T', 'end': 133256042, 'assembly_name': 'GRCh38
', 'colocated_variants': [{'exac_amr_maf': 0.004932, 'exac_sas_maf
': 0.001639, 'exac_nfe_maf': 0.005339, 'end': 133256042, 'eas_maf
': 0, 'exac_adj_allele': 'T', 'exac_eas_maf': 0, 'exac_fin_maf':
0.02601, 'exac_amr_allele': 'T', 'ea_maf': 0.003809, 'exac_oth_all
ele': 'T', 'minor_allele': 'T', 'sas_allele': 'T', 'amr_allele': '
T', 'exac_afr_allele': 'T', 'start': 133256042, 'ea_allele': 'T',
'allele_string': 'C/T', 'exac_eas_allele': 'T', 'afr_maf': 0, 'amr
_maf': 0.0014, 'sas_maf': 0.001, 'eur_allele': 'T', 'exac_fin_alle
le': 'T', 'strand': 1, 'afr_allele': 'T', 'exac_maf': 0.003022, 'm
inor_allele_freq': 0.0026, 'seq_region_name': 9, 'exac_nfe_allele
': 'T', 'exac_adj_maf': 0.004439, 'exac_afr_maf': 0.0005079, 'aa_m
af': 0.0007102, 'exac_sas_allele': 'T', 'eur_maf': 0.0109, 'aa_all
ele': 'T', 'eas_allele': 'T', 'id': 'rs56116432', 'exac_allele': '
T', 'exac_oth_maf': 0.004926}], 'strand': 1, 'id': 'rs56116432', '
start': 133256042}
```


In [13]:
```
print (len(data[0]["transcript_consequences"]))
```

```
3
```

Η πρώτη εγγραφή έχει consequences σε 3 transcripts. Ας πάρουμε το 2ο:

In [14]:
```
print (data[0]["transcript_consequences"][1])
```

```
{'biotype': 'protein_coding', 'gene_id': 'ENSG00000175164', 'sift_
score': 0, 'hgnc_id': 'HGNC:79', 'polyphen_score': 0.997, 'sift_pr
ediction': 'deleterious', 'variant_allele': 'T', 'cds_start': 686,
'codons': 'gGc/gAc', 'transcript_id': 'ENST00000538324', 'cds_end
': 686, 'polyphen_prediction': 'probably_damaging', 'impact': 'MOD
ERATE', 'cdna_end': 711, 'gene_symbol_source': 'HGNC', 'cdna_start
': 711, 'consequence_terms': ['missense_variant'], 'protein_end':
229, 'protein_start': 229, 'strand': -1, 'amino_acids': 'G/D', 'ge
ne_symbol': 'ABO'}
```

Ας δούμε ποιο είναι το consequence αυτού του mutation σε αυτό το transcript σύμφωνα με το polyphen (http://genetics.bwh.harvard.edu/pph2/):

In [15]:
```
print (data[0]["transcript_consequences"][1]["polyphen_predictio
n"])
```

```
probably_damaging
```

POST request για dbSNP variant

Θα κάνουμε τώρα το ίδιο αλλά θα χρησιμοποιήσουμε το API για POST request: http://rest.ensembl.org/documentation/info/vep_id_post (http://rest.ensembl.org/documentation/info/vep_id_post)

In [17]:
```
url = 'http://rest.ensembl.org/vep/human/id'

# δηλώνουμε ότι το αποτέλεσμα θέλουμε να είναι σε JSON φορματ (Acce
pt)
headers = { "Content-Type" : "application/json", "Accept" : "applic
ation/json"}

# Αφού είναι POST βάζουμε τα data ξεχωριστά:
data = { "ids" : ["rs56116432", "COSM476" ] }

# ΠΡΟΣΟΧΗ!!!
# Τα data είναι ένα dictionary το οποίο πρέπει να το μετατρέψουμε σ
ε JSON!
import json
data_json = json.dumps(data)

# Κάνουμε το POST request:
r = requests.post(url, headers=headers, data=data_json)
```

Τσεκάρουμε ότι όλα πήγαν οκ:

In [18]:
```
print (r.ok)
```

```
True
```

παίρνουμε τα data:

In [19]:
```
data = r.json()
print (data)
```

[{'transcript_consequences': [{'biotype': 'processed_transcript',
'gene_id': 'ENSG00000175164', 'cdna_end': 700, 'hgnc_id': 'HGNC:79
', 'gene_symbol_source': 'HGNC', 'cdna_start': 700, 'variant_allel
e': 'T', 'transcript_id': 'ENST00000453660', 'consequence_terms':
['non_coding_transcript_exon_variant', 'non_coding_transcript_vari
ant'], 'impact': 'MODIFIER', 'strand': -1, 'gene_symbol': 'ABO'},
{'biotype': 'protein_coding', 'gene_id': 'ENSG00000175164', 'sift_
score': 0, 'hgnc_id': 'HGNC:79', 'polyphen_score': 0.997, 'sift_pr
ediction': 'deleterious', 'variant_allele': 'T', 'cds_start': 686,
'codons': 'gGc/gAc', 'transcript_id': 'ENST00000538324', 'cds_end
': 686, 'polyphen_prediction': 'probably_damaging', 'impact': 'MOD
ERATE', 'cdna_end': 711, 'gene_symbol_source': 'HGNC', 'cdna_start
': 711, 'consequence_terms': ['missense_variant'], 'protein_end':
229, 'protein_start': 229, 'strand': -1, 'amino_acids': 'G/D', 'ge
ne_symbol': 'ABO'}, {'biotype': 'protein_coding', 'gene_id': 'ENSG
00000175164', 'sift_score': 0, 'hgnc_id': 'HGNC:79', 'polyphen_sco
re': 0.997, 'sift_prediction': 'deleterious', 'variant_allele': 'T
', 'cds_start': 686, 'codons': 'gGc/gAc', 'transcript_id': 'ENST00
000611156', 'cds_end': 686, 'polyphen_prediction': 'probably_damag
ing', 'impact': 'MODERATE', 'cdna_end': 711, 'gene_symbol_source':
'HGNC', 'cdna_start': 711, 'consequence_terms': ['missense_variant
'], 'protein_end': 229, 'protein_start': 229, 'strand': -1, 'amino
_acids': 'G/D', 'gene_symbol': 'ABO'}], 'most_severe_consequence':
'missense_variant', 'input': 'rs56116432', 'seq_region_name': '9',
'allele_string': 'C/T', 'end': 133256042, 'assembly_name': 'GRCh38
', 'colocated_variants': [{'exac_amr_maf': 0.004932, 'exac_sas_maf
': 0.001639, 'exac_nfe_maf': 0.005339, 'end': 133256042, 'eas_maf
': 0, 'exac_adj_allele': 'T', 'exac_eas_maf': 0, 'exac_fin_maf':
0.02601, 'exac_amr_allele': 'T', 'ea_maf': 0.003809, 'exac_oth_all
ele': 'T', 'minor_allele': 'T', 'sas_allele': 'T', 'amr_allele': '
T', 'exac_afr_allele': 'T', 'start': 133256042, 'ea_allele': 'T',
'allele_string': 'C/T', 'exac_eas_allele': 'T', 'afr_maf': 0, 'amr
_maf': 0.0014, 'sas_maf': 0.001, 'eur_allele': 'T', 'exac_fin_alle
le': 'T', 'strand': 1, 'afr_allele': 'T', 'exac_maf': 0.003022, 'm
inor_allele_freq': 0.0026, 'seq_region_name': 9, 'exac_nfe_allele
': 'T', 'exac_adj_maf': 0.004439, 'exac_afr_maf': 0.0005079, 'aa_m
af': 0.0007102, 'exac_sas_allele': 'T', 'eur_maf': 0.0109, 'aa_all
ele': 'T', 'eas_allele': 'T', 'id': 'rs56116432', 'exac_allele': '
T', 'exac_oth_maf': 0.004926}], 'strand': 1, 'id': 'rs56116432', '
start': 133256042}, {'transcript_consequences': [{'biotype': 'prot
ein_coding', 'gene_id': 'ENSG00000281879', 'sift_score': 0, 'hgnc_
id': 'HGNC:79', 'polyphen_score': 0.997, 'sift_prediction': 'delet
erious', 'variant_allele': 'T', 'cds_start': 689, 'codons': 'gGc/g
Ac', 'transcript_id': 'ENST00000626615', 'cds_end': 689, 'polyphen
_prediction': 'probably_damaging', 'impact': 'MODERATE', 'cdna_end
': 714, 'gene_symbol_source': 'HGNC', 'cdna_start': 714, 'conseque
nce_terms': ['missense_variant'], 'protein_end': 230, 'protein_sta
rt': 230, 'strand': -1, 'amino_acids': 'G/D', 'gene_symbol': 'ABO
'}], 'most_severe_consequence': 'missense_variant', 'input': 'rs56
116432', 'seq_region_name': 'CHR_HG2030_PATCH', 'allele_string': '
C/T', 'end': 133256189, 'assembly_name': 'GRCh38', 'colocated_vari
ants': [{'allele_string': 'C/T', 'seq_region_name': 'CHR_HG2030_PA
TCH', 'end': 133256189, 'minor_allele': 'T', 'strand': 1, 'id': 'r
s56116432', 'minor_allele_freq': 0.0026, 'start': 133256189}], 'st
rand': 1, 'id': 'rs56116432', 'start': 133256189}, {'most_severe_c
onsequence': '?', 'allele_string': 'COSMIC_MUTATION', 'input': 'CO
SM476', 'seq_region_name': '7', 'end': 140753336, 'assembly_name':
'GRCh38', 'colocated_variants': [{'allele_string': 'HGMD_MUTATION
', 'seq_region_name': 7, 'phenotype_or_disease': 1, 'strand': 1, '

```
id': 'CM112509', 'end': 140753336, 'start': 140753336}, {'allele_s
tring': 'COSMIC_MUTATION', 'seq_region_name': 7, 'phenotype_or_dis
ease': 1, 'start': 140753336, 'strand': 1, 'id': 'COSM18443', 'end
': 140753336, 'somatic': 1}, {'allele_string': 'COSMIC_MUTATION',
'seq_region_name': 7, 'phenotype_or_disease': 1, 'start': 14075333
6, 'strand': 1, 'id': 'COSM476', 'end': 140753336, 'somatic': 1},
{'allele_string': 'COSMIC_MUTATION', 'seq_region_name': 7, 'phenot
ype_or_disease': 1, 'start': 140753336, 'strand': 1, 'id': 'COSM61
37', 'end': 140753336, 'somatic': 1}], 'strand': 1, 'id': 'COSM476
', 'start': 140753336}]
```

Είναι τα ίδια με πριν:

In [20]:
```
print (data[0]["transcript_consequences"][1]["polyphen_predictio
n"])
```

```
probably_damaging
```

GET request για HGVS variant

Η ENSEMBL μας δίνει ένα API για να πάρουμε πληροφορίες από το Variant Effect Predictor για ένα HGVS variant μέσω GET request: http://rest.ensembl.org/documentation/info/vep_hgvs_get (http://rest.ensembl.org/documentation/info/vep_hgvs_get)

In [21]:
```
url = 'http://rest.ensembl.org/vep/human/hgvs/9:g.22125504G>C?'
headers={ "Content-Type" : "application/json"}

r = requests.get(url, headers=headers)
```

In [22]:
```
print (r.ok)
```

```
True
```

In [23]:
```
data = r.json()
```

In [24]:
```
print (data)
```

[{'transcript_consequences': [{'consequence_terms': ['downstream_g
ene_variant'], 'impact': 'MODIFIER', 'biotype': 'antisense', 'gene
_id': 'ENSG00000240498', 'hgnc_id': 'HGNC:34341', 'distance': 440
7, 'gene_symbol': 'CDKN2B-AS1', 'gene_symbol_source': 'HGNC', 'str
and': 1, 'variant_allele': 'C', 'transcript_id': 'ENST00000422420
'}, {'consequence_terms': ['downstream_gene_variant'], 'impact': '
MODIFIER', 'biotype': 'antisense', 'gene_id': 'ENSG00000240498', '
hgnc_id': 'HGNC:34341', 'distance': 4409, 'gene_symbol': 'CDKN2B-A
S1', 'gene_symbol_source': 'HGNC', 'strand': 1, 'variant_allele':
'C', 'transcript_id': 'ENST00000428597'}, {'consequence_terms': ['
downstream_gene_variant'], 'impact': 'MODIFIER', 'biotype': 'antis
ense', 'gene_id': 'ENSG00000240498', 'hgnc_id': 'HGNC:34341', 'dis
tance': 4932, 'gene_symbol': 'CDKN2B-AS1', 'gene_symbol_source': '
HGNC', 'strand': 1, 'variant_allele': 'C', 'transcript_id': 'ENST0
0000577551'}, {'consequence_terms': ['downstream_gene_variant'], '
impact': 'MODIFIER', 'biotype': 'antisense', 'gene_id': 'ENSG00000
240498', 'hgnc_id': 'HGNC:34341', 'distance': 4858, 'gene_symbol':
'CDKN2B-AS1', 'gene_symbol_source': 'HGNC', 'strand': 1, 'variant_
allele': 'C', 'transcript_id': 'ENST00000580576'}, {'consequence_t
erms': ['downstream_gene_variant'], 'impact': 'MODIFIER', 'biotype
': 'antisense', 'gene_id': 'ENSG00000240498', 'hgnc_id': 'HGNC:343
41', 'distance': 4932, 'gene_symbol': 'CDKN2B-AS1', 'gene_symbol_s
ource': 'HGNC', 'strand': 1, 'variant_allele': 'C', 'transcript_id
': 'ENST00000581051'}, {'consequence_terms': ['downstream_gene_var
iant'], 'impact': 'MODIFIER', 'biotype': 'antisense', 'gene_id': '
ENSG00000240498', 'hgnc_id': 'HGNC:34341', 'distance': 4932, 'gene
_symbol': 'CDKN2B-AS1', 'gene_symbol_source': 'HGNC', 'strand': 1,
'variant_allele': 'C', 'transcript_id': 'ENST00000582072'}, {'cons
equence_terms': ['downstream_gene_variant'], 'impact': 'MODIFIER',
'biotype': 'antisense', 'gene_id': 'ENSG00000240498', 'hgnc_id': '
HGNC:34341', 'distance': 4932, 'gene_symbol': 'CDKN2B-AS1', 'gene_
symbol_source': 'HGNC', 'strand': 1, 'variant_allele': 'C', 'trans
cript_id': 'ENST00000584020'}, {'consequence_terms': ['downstream_
gene_variant'], 'impact': 'MODIFIER', 'biotype': 'antisense', 'gen
e_id': 'ENSG00000240498', 'hgnc_id': 'HGNC:34341', 'distance': 493
2, 'gene_symbol': 'CDKN2B-AS1', 'gene_symbol_source': 'HGNC', 'str
and': 1, 'variant_allele': 'C', 'transcript_id': 'ENST00000584637
'}, {'consequence_terms': ['downstream_gene_variant'], 'impact': '
MODIFIER', 'biotype': 'antisense', 'gene_id': 'ENSG00000240498', '
hgnc_id': 'HGNC:34341', 'distance': 4932, 'gene_symbol': 'CDKN2B-A
S1', 'gene_symbol_source': 'HGNC', 'strand': 1, 'variant_allele':
'C', 'transcript_id': 'ENST00000584816'}, {'consequence_terms': ['
downstream_gene_variant'], 'impact': 'MODIFIER', 'biotype': 'antis
ense', 'gene_id': 'ENSG00000240498', 'hgnc_id': 'HGNC:34341', 'dis
tance': 4960, 'gene_symbol': 'CDKN2B-AS1', 'gene_symbol_source': '
HGNC', 'strand': 1, 'variant_allele': 'C', 'transcript_id': 'ENST0
0000585267'}], 'most_severe_consequence': 'downstream_gene_variant
', 'input': '9:g.22125504G>C', 'seq_region_name': '9', 'allele_str
ing': 'G/C', 'end': 22125504, 'assembly_name': 'GRCh38', 'colocate
d_variants': [{'allele_string': 'G/C', 'eur_maf': 0.4722, 'amr_maf
': 0.4553, 'sas_maf': 0.4908, 'eur_allele': 'C', 'end': 22125504,
'eas_maf': 0.5367, 'afr_allele': 'C', 'minor_allele': 'C', 'seq_re
gion_name': 9, 'minor_allele_freq': 0.4181, 'strand': 1, 'pubmed':
'24262325,19501493,22042884,21860704,21149552,20159871,19474294,21
894447,21971053,21804106,20502693,22199011,18224312,22400124,18533
027,18852197,21297524,22403240,22856518,23963167,19343170,2038674
0,21400687,24728607,20017983,24573017,24607648,20549515,22144573,2
2623978,22029572,18362232,19173706,19214202,26252781,20435227,2160
6135,19924713,17554300,19955471,19956433,25717410,24098343,1878030

```
2,18675980,19475673,20231156,20858905,21152093,21698238,24906238,1
7634449,18979498,19164808,19207022,19750184,20098575,20981302,2124
2481,21369780,22295058,22848412,25617895,23729007,18469204,2060502
3,21372283,22429504,26483964,23870195,18704761,23587283,24926413,1
9463184,24676469,21424681,20175863,22505696,19559344,19578366,2314
2796,19171343,24246088,18987759,19819472,19926059,21375403,2138535
5,21705410,24777168,25105296,19888323,23454037,18264662,18599798,1
8652946,18654002,18925945,18957718,19135198,19319159,19329499,1954
8844,19664850,19709660,19885677,19956784,20031540,20031580,2003160
6,20230275,20335276,20395598,2040077', 'sas_allele': 'C', 'phenoty
pe_or_disease': 1, 'amr_allele': 'C', 'afr_maf': 0.2133, 'eas_alle
le': 'C', 'id': 'rs1333049', 'start': 22125504}], 'strand': 1, 'id
': '9:g.22125504G>C', 'start': 22125504}]
```

In [25]:
```
print (len(data))
```

```
1
```

Επέστρεψε μόνο μία εγγρφή.

In [26]:
```
data[0].keys()
```

Out[26]:
```
dict_keys(['transcript_consequences', 'most_severe_consequence', '
input', 'seq_region_name', 'allele_string', 'end', 'assembly_name
', 'colocated_variants', 'strand', 'id', 'start'])
```

Ας δούμε τα transcript consequences:

In [27]:
```
data[0]["transcript_consequences"]
```

Out[27]:

```
[{'biotype': 'antisense',
  'consequence_terms': ['downstream_gene_variant'],
  'distance': 4407,
  'gene_id': 'ENSG00000240498',
  'gene_symbol': 'CDKN2B-AS1',
  'gene_symbol_source': 'HGNC',
  'hgnc_id': 'HGNC:34341',
  'impact': 'MODIFIER',
  'strand': 1,
  'transcript_id': 'ENST00000422420',
  'variant_allele': 'C'},
 {'biotype': 'antisense',
  'consequence_terms': ['downstream_gene_variant'],
  'distance': 4409,
  'gene_id': 'ENSG00000240498',
  'gene_symbol': 'CDKN2B-AS1',
  'gene_symbol_source': 'HGNC',
  'hgnc_id': 'HGNC:34341',
  'impact': 'MODIFIER',
  'strand': 1,
  'transcript_id': 'ENST00000428597',
  'variant_allele': 'C'},
 {'biotype': 'antisense',
  'consequence_terms': ['downstream_gene_variant'],
  'distance': 4932,
  'gene_id': 'ENSG00000240498',
  'gene_symbol': 'CDKN2B-AS1',
  'gene_symbol_source': 'HGNC',
  'hgnc_id': 'HGNC:34341',
  'impact': 'MODIFIER',
  'strand': 1,
  'transcript_id': 'ENST00000577551',
  'variant_allele': 'C'},
 {'biotype': 'antisense',
  'consequence_terms': ['downstream_gene_variant'],
  'distance': 4858,
  'gene_id': 'ENSG00000240498',
  'gene_symbol': 'CDKN2B-AS1',
  'gene_symbol_source': 'HGNC',
  'hgnc_id': 'HGNC:34341',
  'impact': 'MODIFIER',
  'strand': 1,
  'transcript_id': 'ENST00000580576',
  'variant_allele': 'C'},
 {'biotype': 'antisense',
  'consequence_terms': ['downstream_gene_variant'],
  'distance': 4932,
  'gene_id': 'ENSG00000240498',
  'gene_symbol': 'CDKN2B-AS1',
  'gene_symbol_source': 'HGNC',
  'hgnc_id': 'HGNC:34341',
  'impact': 'MODIFIER',
  'strand': 1,
  'transcript_id': 'ENST00000581051',
  'variant_allele': 'C'},
 {'biotype': 'antisense',
  'consequence_terms': ['downstream_gene_variant'],
  'distance': 4932,
  'gene_id': 'ENSG00000240498',
```

```
 'gene_symbol': 'CDKN2B-AS1',
 'gene_symbol_source': 'HGNC',
 'hgnc_id': 'HGNC:34341',
 'impact': 'MODIFIER',
 'strand': 1,
 'transcript_id': 'ENST00000582072',
 'variant_allele': 'C'},
{'biotype': 'antisense',
 'consequence_terms': ['downstream_gene_variant'],
 'distance': 4932,
 'gene_id': 'ENSG00000240498',
 'gene_symbol': 'CDKN2B-AS1',
 'gene_symbol_source': 'HGNC',
 'hgnc_id': 'HGNC:34341',
 'impact': 'MODIFIER',
 'strand': 1,
 'transcript_id': 'ENST00000584020',
 'variant_allele': 'C'},
{'biotype': 'antisense',
 'consequence_terms': ['downstream_gene_variant'],
 'distance': 4932,
 'gene_id': 'ENSG00000240498',
 'gene_symbol': 'CDKN2B-AS1',
 'gene_symbol_source': 'HGNC',
 'hgnc_id': 'HGNC:34341',
 'impact': 'MODIFIER',
 'strand': 1,
 'transcript_id': 'ENST00000584637',
 'variant_allele': 'C'},
{'biotype': 'antisense',
 'consequence_terms': ['downstream_gene_variant'],
 'distance': 4932,
 'gene_id': 'ENSG00000240498',
 'gene_symbol': 'CDKN2B-AS1',
 'gene_symbol_source': 'HGNC',
 'hgnc_id': 'HGNC:34341',
 'impact': 'MODIFIER',
 'strand': 1,
 'transcript_id': 'ENST00000584816',
 'variant_allele': 'C'},
{'biotype': 'antisense',
 'consequence_terms': ['downstream_gene_variant'],
 'distance': 4960,
 'gene_id': 'ENSG00000240498',
 'gene_symbol': 'CDKN2B-AS1',
 'gene_symbol_source': 'HGNC',
 'hgnc_id': 'HGNC:34341',
 'impact': 'MODIFIER',
 'strand': 1,
 'transcript_id': 'ENST00000585267',
 'variant_allele': 'C'}]
```


POST request για HGVS variant

Μπορούμε να κάνουμε το ίδιο, κάνοντας ένα POST request: http://rest.ensembl.org/documentation/info/vep_hgvs_post (http://rest.ensembl.org/documentation/info/vep_hgvs_post)

In [28]:
```
url = 'http://rest.ensembl.org/vep/human/hgvs'
headers = { "Content-Type" : "application/json", "Accept" : "applic
ation/json"}
data = { "hgvs_notations" : ["9:g.22125504G>C"] } # Παρατηρείστε ότ
ι αυτό είναι μία λίστα.
                                                  # Μπορούμε να βάλ
ουμε (σχεδόν) όσα HGVS variants θέλουμε

data_json = json.dumps(data)

r = requests.post(url, headers=headers, data=data_json)
```

In [29]:
```
print (r.ok)
```

```
True
```

In [30]:
```
returned_data = r.json()
print (returned_data)
```

[{'transcript_consequences': [{'consequence_terms': ['downstream_g
ene_variant'], 'impact': 'MODIFIER', 'biotype': 'antisense', 'gene
_id': 'ENSG00000240498', 'hgnc_id': 'HGNC:34341', 'distance': 440
7, 'gene_symbol': 'CDKN2B-AS1', 'gene_symbol_source': 'HGNC', 'str
and': 1, 'variant_allele': 'C', 'transcript_id': 'ENST00000422420
'}, {'consequence_terms': ['downstream_gene_variant'], 'impact': '
MODIFIER', 'biotype': 'antisense', 'gene_id': 'ENSG00000240498', '
hgnc_id': 'HGNC:34341', 'distance': 4409, 'gene_symbol': 'CDKN2B-A
S1', 'gene_symbol_source': 'HGNC', 'strand': 1, 'variant_allele':
'C', 'transcript_id': 'ENST00000428597'}, {'consequence_terms': ['
downstream_gene_variant'], 'impact': 'MODIFIER', 'biotype': 'antis
ense', 'gene_id': 'ENSG00000240498', 'hgnc_id': 'HGNC:34341', 'dis
tance': 4932, 'gene_symbol': 'CDKN2B-AS1', 'gene_symbol_source': '
HGNC', 'strand': 1, 'variant_allele': 'C', 'transcript_id': 'ENST0
0000577551'}, {'consequence_terms': ['downstream_gene_variant'], '
impact': 'MODIFIER', 'biotype': 'antisense', 'gene_id': 'ENSG00000
240498', 'hgnc_id': 'HGNC:34341', 'distance': 4858, 'gene_symbol':
'CDKN2B-AS1', 'gene_symbol_source': 'HGNC', 'strand': 1, 'variant_
allele': 'C', 'transcript_id': 'ENST00000580576'}, {'consequence_t
erms': ['downstream_gene_variant'], 'impact': 'MODIFIER', 'biotype
': 'antisense', 'gene_id': 'ENSG00000240498', 'hgnc_id': 'HGNC:343
41', 'distance': 4932, 'gene_symbol': 'CDKN2B-AS1', 'gene_symbol_s
ource': 'HGNC', 'strand': 1, 'variant_allele': 'C', 'transcript_id
': 'ENST00000581051'}, {'consequence_terms': ['downstream_gene_var
iant'], 'impact': 'MODIFIER', 'biotype': 'antisense', 'gene_id': '
ENSG00000240498', 'hgnc_id': 'HGNC:34341', 'distance': 4932, 'gene
_symbol': 'CDKN2B-AS1', 'gene_symbol_source': 'HGNC', 'strand': 1,
'variant_allele': 'C', 'transcript_id': 'ENST00000582072'}, {'cons
equence_terms': ['downstream_gene_variant'], 'impact': 'MODIFIER',
'biotype': 'antisense', 'gene_id': 'ENSG00000240498', 'hgnc_id': '
HGNC:34341', 'distance': 4932, 'gene_symbol': 'CDKN2B-AS1', 'gene_
symbol_source': 'HGNC', 'strand': 1, 'variant_allele': 'C', 'trans
cript_id': 'ENST00000584020'}, {'consequence_terms': ['downstream_
gene_variant'], 'impact': 'MODIFIER', 'biotype': 'antisense', 'gen
e_id': 'ENSG00000240498', 'hgnc_id': 'HGNC:34341', 'distance': 493
2, 'gene_symbol': 'CDKN2B-AS1', 'gene_symbol_source': 'HGNC', 'str
and': 1, 'variant_allele': 'C', 'transcript_id': 'ENST00000584637
'}, {'consequence_terms': ['downstream_gene_variant'], 'impact': '
MODIFIER', 'biotype': 'antisense', 'gene_id': 'ENSG00000240498', '
hgnc_id': 'HGNC:34341', 'distance': 4932, 'gene_symbol': 'CDKN2B-A
S1', 'gene_symbol_source': 'HGNC', 'strand': 1, 'variant_allele':
'C', 'transcript_id': 'ENST00000584816'}, {'consequence_terms': ['
downstream_gene_variant'], 'impact': 'MODIFIER', 'biotype': 'antis
ense', 'gene_id': 'ENSG00000240498', 'hgnc_id': 'HGNC:34341', 'dis
tance': 4960, 'gene_symbol': 'CDKN2B-AS1', 'gene_symbol_source': '
HGNC', 'strand': 1, 'variant_allele': 'C', 'transcript_id': 'ENST0
0000585267'}], 'most_severe_consequence': 'downstream_gene_variant
', 'input': '9:g.22125504G>C', 'seq_region_name': '9', 'allele_str
ing': 'G/C', 'end': 22125504, 'assembly_name': 'GRCh38', 'colocate
d_variants': [{'allele_string': 'G/C', 'eur_maf': 0.4722, 'amr_maf
': 0.4553, 'sas_maf': 0.4908, 'eur_allele': 'C', 'end': 22125504,
'eas_maf': 0.5367, 'afr_allele': 'C', 'minor_allele': 'C', 'seq_re
gion_name': 9, 'minor_allele_freq': 0.4181, 'strand': 1, 'pubmed':
'24262325,19501493,22042884,21860704,21149552,20159871,19474294,21
894447,21971053,21804106,20502693,22199011,18224312,22400124,18533
027,18852197,21297524,22403240,22856518,23963167,19343170,2038674
0,21400687,24728607,20017983,24573017,24607648,20549515,22144573,2
2623978,22029572,18362232,19173706,19214202,26252781,20435227,2160
6135,19924713,17554300,19955471,19956433,25717410,24098343,1878030

```
2,18675980,19475673,20231156,20858905,21152093,21698238,24906238,1
7634449,18979498,19164808,19207022,19750184,20098575,20981302,2124
2481,21369780,22295058,22848412,25617895,23729007,18469204,2060502
3,21372283,22429504,26483964,23870195,18704761,23587283,24926413,1
9463184,24676469,21424681,20175863,22505696,19559344,19578366,2314
2796,19171343,24246088,18987759,19819472,19926059,21375403,2138535
5,21705410,24777168,25105296,19888323,23454037,18264662,18599798,1
8652946,18654002,18925945,18957718,19135198,19319159,19329499,1954
8844,19664850,19709660,19885677,19956784,20031540,20031580,2003160
6,20230275,20335276,20395598,2040077', 'sas_allele': 'C', 'phenoty
pe_or_disease': 1, 'amr_allele': 'C', 'afr_maf': 0.2133, 'eas_alle
le': 'C', 'id': 'rs1333049', 'start': 22125504}], 'strand': 1, 'id
': '9:g.22125504C>C', 'start': 22125504}]
```

Mutalyzer

Το mutalyzer (https://mutalyzer.nl/) είναι ένα service με το οποίο μπορούμε να ελέγχουμε και να κάνουμε μετατροπές σε variants οι οποίοι περιγράφονται σε HGVS. Για παράδειγμα μέσω το numberConversion μπορούμε να βρούμε το genomic location ενός variant:

In [2]:
```
r = requests.get(
  'https://mutalyzer.nl/json/numberConversion',
  {
  'build': 'hg38',
  'variant': 'NM_001276506.1:c.204C>T'
  })
data = r.json()
print (data)
```

```
['NC_000011.10:g.112088901C>T']
```

Variation Reporter

Το Variation Reporter (https://www.ncbi.nlm.nih.gov/variation/tools/reporter) είναι το αντίστοιχο με το VEP service του NCBI. Μπορουμε να το χρησιμοποιήσουμε ως εξής:

In [4]:
```
url='https://www.ncbi.nlm.nih.gov/projects/SNP/VariantAnalyzer/var_
rep.cgi'
var = 'NM_001276506.1:c.204C>T'
r = requests.post(url, {'annot1': var, })
print (r.ok)
data = r.text
print (data)
```

```
True
Submitted: JSID_01_1003310_130.14.18.6_9000__variant_analyzer
.
## URL: https://www.ncbi.nlm.nih.gov/variation/tools/reporter/JSID
_01_1003310_130.14.18.6_9000__variant_analyzer
## Submitted time:      01/26/2018 11:29:20
## Report generated time:       01/26/2018 11:29:22
## Summary report for
## Summary of Submitted Data
## Number of variant locations: 1
## Number of variant alleles:   1
## Summary of Data Report:
## Number of unique NCBI Ids found:     1
## Number of novel alleles at known locations:  0
## Number of novel alleles at novel locations:  0
## Total number of novel locations:     0
## Variant alleles with clinical information:   1
## Variant alleles with molecular consequence:  1
## Total number of rows in the report:  13
## Assembly:    GRCh37.p13

# Submitted ID  Submitted Loc   Cytoband        Mapped Loc      NC
BI ID   Allele  GMAF    Gene ID Gene Symbol     Hgvs_g  Origin  Hg
vs_g (RefSeqGene)       Hgvs_c  Hgvs_p  Consequences    OnTestPane
l       ClinVar Accession       Clinical Source ID      Clinical S
ignificance     Clinical Evidence       Clinical Review Phenotype
ID      Phenotype Description   PMIDs   Number Of Submissions   Su
spected False Positive  GWAS Association        Has Genotype?   On
Genotyping Kit? Exception       Novel allele?
NM_001276506.1:c.204C>T NM_001276506.1: 288     11q23.1 NP_002993.
1:68    rs9919552       NC_000011.9:g.111959625C>T      T: 0.1134
        6392    SDHD    NC_000011.9:g.111959625C>T      germline
        NG_012337.3:g.7055C>T   NM_001276506.1:c.204C>T NP_0012634
35.1:p.Ser68=   SO:0001819 (synonymous_codon)           RCV0000377
27.5; RCV000162450.1; RCV000265027.1    Illumina Clinical Services
Laboratory,Illumina:839653      Conflicting interpretations of pat
hogenicity; Benign; Benign      clinical testing; clinical testin
g; clinical testing     conf; single; single    MedGen:CN169374; M
edGen:C0027672,SNOMED CT:699346009; MedGen:C0031511,OMIM:171300 no
t specified; Hereditary cancer-predisposing syndrome; Pheochromocy
toma    24033266;25741868       34                      Yes     Ye
s
NM_001276506.1:c.204C>T NM_001276506.1: 288     11q23.1 NT_033899.
8:15522041      rs9919552       NC_000011.9:g.111959625C>T      T:
0.1134  6392    SDHD    NC_000011.9:g.111959625C>T      germline
        NG_012337.3:g.7055C>T   NM_001276506.1:c.204C>T NP_0012634
35.1:p.Ser68=   SO:0001819 (synonymous_codon)           RCV0000377
27.5; RCV000162450.1; RCV000265027.1    Illumina Clinical Services
Laboratory,Illumina:839653      Conflicting interpretations of pat
hogenicity; Benign; Benign      clinical testing; clinical testin
g; clinical testing     conf; single; single    MedGen:CN169374; M
edGen:C0027672,SNOMED CT:699346009; MedGen:C0031511,OMIM:171300 no
t specified; Hereditary cancer-predisposing syndrome; Pheochromocy
toma    24033266;25741868       34                      Yes     Ye
s
NM_001276506.1:c.204C>T NM_001276506.1: 288     11q23.1 NC_000011.
9:111959625     rs9919552       NC_000011.9:g.111959625C>T      T:
0.1134  6392    SDHD    NC_000011.9:g.111959625C>T      germline
        NG_012337.3:g.7055C>T   NM_001276506.1:c.204C>T NP_0012634
```

```
35.1:p.Ser68=   SO:0001819 (synonymous_codon)           RCV0000377
27.5; RCV000162450.1; RCV000265027.1    Illumina Clinical Services
Laboratory,Illumina:839653      Conflicting interpretations of pat
hogenicity; Benign; Benign      clinical testing; clinical testin
g; clinical testing     conf; single; single    MedGen:CN169374; M
edGen:C0027672,SNOMED CT:699346009; MedGen:C0031511,OMIM:171300 no
t specified; Hereditary cancer-predisposing syndrome; Pheochromocy
toma    24033266;25741868       34                      Yes     Ye
s
NM_001276506.1:c.204C>T NM_001276506.1: 288     11q23.1 NM_003002.
3:288   rs9919552       NC_000011.9:g.111959625C>T      T: 0.1134
        6392    SDHD    NC_000011.9:g.111959625C>T      germline
        NG_012337.3:g.7055C>T   NM_001276506.1:c.204C>T NP_0012634
35.1:p.Ser68=   SO:0001819 (synonymous_codon)           RCV0000377
27.5; RCV000162450.1; RCV000265027.1    Illumina Clinical Services
Laboratory,Illumina:839653      Conflicting interpretations of pat
hogenicity; Benign; Benign      clinical testing; clinical testin
g; clinical testing     conf; single; single    MedGen:CN169374; M
edGen:C0027672,SNOMED CT:699346009; MedGen:C0031511,OMIM:171300 no
t specified; Hereditary cancer-predisposing syndrome; Pheochromocy
toma    24033266;25741868       34                      Yes     Ye
s
NM_001276506.1:c.204C>T NM_001276506.1: 288     11q23.1 NM_0012765
04.1:171        rs9919552       NC_000011.9:g.111959625C>T      T:
0.1134  6392    SDHD    NC_000011.9:g.111959625C>T      germline
        NG_012337.3:g.7055C>T   NM_001276506.1:c.204C>T NP_0012634
35.1:p.Ser68=   SO:0001819 (synonymous_codon)           RCV0000377
27.5; RCV000162450.1; RCV000265027.1    Illumina Clinical Services
Laboratory,Illumina:839653      Conflicting interpretations of pat
hogenicity; Benign; Benign      clinical testing; clinical testin
g; clinical testing     conf; single; single    MedGen:CN169374; M
edGen:C0027672,SNOMED CT:699346009; MedGen:C0031511,OMIM:171300 no
t specified; Hereditary cancer-predisposing syndrome; Pheochromocy
toma    24033266;25741868       34                      Yes     Ye
s
NM_001276506.1:c.204C>T NM_001276506.1: 288     11q23.1 NP_0012634
33.1:29 rs9919552       NC_000011.9:g.111959625C>T      T: 0.1134
        6392    SDHD    NC_000011.9:g.111959625C>T      germline
        NG_012337.3:g.7055C>T   NM_001276506.1:c.204C>T NP_0012634
35.1:p.Ser68=   SO:0001819 (synonymous_codon)           RCV0000377
27.5; RCV000162450.1; RCV000265027.1    Illumina Clinical Services
Laboratory,Illumina:839653      Conflicting interpretations of pat
hogenicity; Benign; Benign      clinical testing; clinical testin
g; clinical testing     conf; single; single    MedGen:CN169374; M
edGen:C0027672,SNOMED CT:699346009; MedGen:C0031511,OMIM:171300 no
t specified; Hereditary cancer-predisposing syndrome; Pheochromocy
toma    24033266;25741868       34                      Yes     Ye
s
NM_001276506.1:c.204C>T NM_001276506.1: 288     11q23.1 NM_0012765
06.1:288        rs9919552       NC_000011.9:g.111959625C>T      T:
0.1134  6392    SDHD    NC_000011.9:g.111959625C>T      germline
        NG_012337.3:g.7055C>T   NM_001276506.1:c.204C>T NP_0012634
35.1:p.Ser68=   SO:0001819 (synonymous_codon)           RCV0000377
27.5; RCV000162450.1; RCV000265027.1    Illumina Clinical Services
Laboratory,Illumina:839653      Conflicting interpretations of pat
hogenicity; Benign; Benign      clinical testing; clinical testin
g; clinical testing     conf; single; single    MedGen:CN169374; M
edGen:C0027672,SNOMED CT:699346009; MedGen:C0031511,OMIM:171300 no
t specified; Hereditary cancer-predisposing syndrome; Pheochromocy
toma    24033266;25741868       34                      Yes     Ye
```

```
s
NM_001276506.1:c.204C>T NM_001276506.1: 288     11q23.1 NP_0012634
35.1:68 rs9919552       NC_000011.9:g.111959625C>T      T: 0.1134
        6392    SDHD    NC_000011.9:g.111959625C>T      germline
        NG_012337.3:g.7055C>T   NM_001276506.1:c.204C>T NP_0012634
35.1:p.Ser68=   SO:0001819 (synonymous_codon)           RCV0000377
27.5; RCV000162450.1; RCV000265027.1    Illumina Clinical Services
Laboratory,Illumina:839653      Conflicting interpretations of pat
hogenicity; Benign; Benign      clinical testing; clinical testin
g; clinical testing     conf; single; single    MedGen:CN169374; M
edGen:C0027672,SNOMED CT:699346009; MedGen:C0031511,OMIM:171300 no
t specified; Hereditary cancer-predisposing syndrome; Pheochromocy
toma    24033266;25741868       34                      Yes     Ye
s
NM_001276506.1:c.204C>T NM_001276506.1: 288     11q23.1 NR_077060.
1:288   rs9919552       NC_000011.9:g.111959625C>T      T: 0.1134
        6392    SDHD    NC_000011.9:g.111959625C>T      germline
        NG_012337.3:g.7055C>T   NM_001276506.1:c.204C>T NP_0012634
35.1:p.Ser68=   SO:0001819 (synonymous_codon)           RCV0000377
27.5; RCV000162450.1; RCV000265027.1    Illumina Clinical Services
Laboratory,Illumina:839653      Conflicting interpretations of pat
hogenicity; Benign; Benign      clinical testing; clinical testin
g; clinical testing     conf; single; single    MedGen:CN169374; M
edGen:C0027672,SNOMED CT:699346009; MedGen:C0031511,OMIM:171300 no
t specified; Hereditary cancer-predisposing syndrome; Pheochromocy
toma    24033266;25741868       34                      Yes     Ye
s
NM_001276506.1:c.204C>T NM_001276506.1: 288     11q23.1 NG_033145.
1:2898  rs9919552       NC_000011.9:g.111959625C>T      T: 0.1134
        6392    SDHD    NC_000011.9:g.111959625C>T      germline
        NG_012337.3:g.7055C>T   NM_001276506.1:c.204C>T NP_0012634
35.1:p.Ser68=   SO:0001819 (synonymous_codon)           RCV0000377
27.5; RCV000162450.1; RCV000265027.1    Illumina Clinical Services
Laboratory,Illumina:839653      Conflicting interpretations of pat
hogenicity; Benign; Benign      clinical testing; clinical testin
g; clinical testing     conf; single; single    MedGen:CN169374; M
edGen:C0027672,SNOMED CT:699346009; MedGen:C0031511,OMIM:171300 no
t specified; Hereditary cancer-predisposing syndrome; Pheochromocy
toma    24033266;25741868       34                      Yes     Ye
s
NM_001276506.1:c.204C>T NM_001276506.1: 288     11q23.1 XM_0052716
44.1:283        rs9919552       NC_000011.9:g.111959625C>T      T:
0.1134  6392    SDHD    NC_000011.9:g.111959625C>T      germline
        NG_012337.3:g.7055C>T   NM_001276506.1:c.204C>T NP_0012634
35.1:p.Ser68=   SO:0001819 (synonymous_codon)           RCV0000377
27.5; RCV000162450.1; RCV000265027.1    Illumina Clinical Services
Laboratory,Illumina:839653      Conflicting interpretations of pat
hogenicity; Benign; Benign      clinical testing; clinical testin
g; clinical testing     conf; single; single    MedGen:CN169374; M
edGen:C0027672,SNOMED CT:699346009; MedGen:C0031511,OMIM:171300 no
t specified; Hereditary cancer-predisposing syndrome; Pheochromocy
toma    24033266;25741868       34                      Yes     Ye
s
NM_001276506.1:c.204C>T NM_001276506.1: 288     11q23.1 XP_0052717
01.1:68 rs9919552       NC_000011.9:g.111959625C>T      T: 0.1134
        6392    SDHD    NC_000011.9:g.111959625C>T      germline
        NG_012337.3:g.7055C>T   NM_001276506.1:c.204C>T NP_0012634
35.1:p.Ser68=   SO:0001819 (synonymous_codon)           RCV0000377
27.5; RCV000162450.1; RCV000265027.1    Illumina Clinical Services
Laboratory,Illumina:839653      Conflicting interpretations of pat
```

```
hogenicity; Benign; Benign      clinical testing; clinical testin
g; clinical testing     conf; single; single    MedGen:CN169374; M
edGen:C0027672,SNOMED CT:699346009; MedGen:C0031511,OMIM:171300 no
t specified; Hereditary cancer-predisposing syndrome; Pheochromocy
toma    24033266;25741868       34                      Yes     Ye
s
NM_001276506.1:c.204C>T NM_001276506.1: 288     11q23.1 NG_012337.
3:7055  rs9919552       NC_000011.9:g.111959625C>T      T: 0.1134
        6392    SDHD    NC_000011.9:g.111959625C>T      germline
        NG_012337.3:g.7055C>T   NM_001276506.1:c.204C>T NP_0012634
35.1:p.Ser68=   SO:0001819 (synonymous_codon)           RCV0000377
27.5; RCV000162450.1; RCV000265027.1    Illumina Clinical Services
Laboratory,Illumina:839653      Conflicting interpretations of pat
hogenicity; Benign; Benign      clinical testing; clinical testin
g; clinical testing     conf; single; single    MedGen:CN169374; M
edGen:C0027672,SNOMED CT:699346009; MedGen:C0031511,OMIM:171300 no
t specified; Hereditary cancer-predisposing syndrome; Pheochromocy
toma    24033266;25741868       34                      Yes     Ye
s
```

Παρατηρούμε ότι τα δεδομένα που επιστρέφει δεν είναι σε JSON αλλά tabular. Μπορούμε να τα "φροτώσουμε" σε ένα pandas DataFrame!

In [6]:
```
import pandas as pd
from io import StringIO

# Remove non tabular data
data = [x for x in data.split('\n') if x[:2] not in ['Su', '.', '##
']]
data = '\n'.join(data)

#Create a "virtual" file-like structure that contains the string.
data_f = StringIO(data)

#data_f is a file. Yet it is not stored in the disk.
#Now we can pass it to pandas!
df = pd.read_csv(data_f, sep='\t')
df
```

Out[6]:

	# Submitted ID	Submitted Loc	Cytoband	Mapped Loc	NCBI
0	NM_001276506.1:c.204C>T	NM_001276506.1:288	11q23.1	NP_002993.1:68	rs99195
1	NM_001276506.1:c.204C>T	NM_001276506.1:288	11q23.1	NT_033899.8:15522041	rs99195
2	NM_001276506.1:c.204C>T	NM_001276506.1:288	11q23.1	NC_000011.9:111959625	rs99195
3	NM_001276506.1:c.204C>T	NM_001276506.1:288	11q23.1	NM_003002.3:288	rs99195
4	NM_001276506.1:c.204C>T	NM_001276506.1:288	11q23.1	NM_001276504.1:171	rs99195
5	NM_001276506.1:c.204C>T	NM_001276506.1:288	11q23.1	NP_001263433.1:29	rs99195

	# Submitted ID	Submitted Loc	Cytoband	Mapped Loc	NCBI
6	NM_001276506.1:c.204C>T	NM_001276506.1:288	11q23.1	NM_001276506.1:288	rs99195
7	NM_001276506.1:c.204C>T	NM_001276506.1:288	11q23.1	NP_001263435.1:68	rs99195
8	NM_001276506.1:c.204C>T	NM_001276506.1:288	11q23.1	NR_077060.1:288	rs99195
9	NM_001276506.1:c.204C>T	NM_001276506.1:	11q23.1	NG_033145.1:2898	rs99195

Assembly Converter

Με το Assembly Converter (https://rest.ensembl.org/documentation/info/assembly_map) μπορούμε να μετατρέψουμε μία γενομική θέση από ένα assembly σε ένα άλλο. Για παράδειγμα έστω ότι έχουμε τη θέση 111959625 στο χρωμόσωμα 11 στο assembly GRCh37 και θέλουμε να το μετατρέψουμε στο GRCh38 assembly:

In [8]:
```
location = 111959625
chromosome = 11

url = 'https://rest.ensembl.org/map/human/GRCh37/{chr}:{pos}..{pos}:1/GRCh38'

this_url = url.format(pos=location, chr=chromosome)
headers = {'content-type':'application/json'}

r = requests.get(this_url, headers=headers)

print (r.ok)

data = r.json()
print (data)
```

```
True
{'mappings': [{'original': {'seq_region_name': '11', 'strand': 1,
'coord_system': 'chromosome', 'end': 111959625, 'start': 11195962
5, 'assembly': 'GRCh37'}, 'mapped': {'seq_region_name': '11', 'str
and': 1, 'coord_system': 'chromosome', 'end': 112088901, 'start':
112088901, 'assembly': 'GRCh38'}}]}
```

Το νέο location είναι το:

In [9]:
```
print (data['mappings'][0]['mapped']['start'])
```

```
112088901
```

Biopython / Entrez

Η ENTREZ (https://www.ncbi.nlm.nih.gov/gquery/) είναι μία ομογενοποιημένη βάση δεδομένων από το NCBI (Το αμερικάνικο EMBL). Δυστυχώς το API της ENTREZ δεν είναι τόσο φιλικό όσο της ENSEMBL. Το documentation (https://www.ncbi.nlm.nih.gov/books/NBK25501/) είναι διάσπαρτο και δεν υπάρχουν πολλά παραδείγματα (Και για αυτά που υπάρχουν είναι σε perl (https://www.ncbi.nlm.nih.gov/books/NBK25498/) 😱).

Η biopython (http://biopython.org) είναι το πιο γνωστό και σημαντικό πακέτο της python για πρόσβαση σε γενετική πληροφορία αλλά και επεξεργασία. Από το documentation (http://biopython.org/DIST/docs/tutorial/Tutorial.html) φαίνεται ότι πρόκειται στην ουσία για ένα ολοκληρωμένο περιβάλλον το οποίο δεν προλαβαίνουμε να καλύψουμε όλο. Εδώ θα δούμε πως μπορούμε να χρησιμοποιήσουμε την biopython για να έχουμε πρόσβαση στο NCBI μέσω του Entrez.

Για να εγκαταστήσουμε τη biopython τρέχουμε:

```
pip install biopython
```

ΠΡΟΣΟΧΗ: Σας προτείνω να ρίξετε τουλάχιστον μία ματιά στο υπέροχο tutorial της biopython: http://biopython.org/DIST/docs/tutorial/Tutorial.html (http://biopython.org/DIST/docs/tutorial/Tutorial.html)

Δύσκολα δεν θα βρείτε κάτι σχετικό με την έρευνα που κάνετε!

Αρχικά κάνουμε import τις βιβλιοθήκες:

In [5]:
```
from Bio import Entrez
import pandas as pd

# For some reason Entrez needs to know who are you
# but you don't have to be that honest..
Entrez.email = 'anonymous@gmail.com'
```

Ας δούμε σε ποιες βάσεις δεδομένων παρέχει πρόσβαση η Entrez:

In [6]:
```
handle = Entrez.einfo()
record = Entrez.read(handle)
```

In [7]:
```
pd.DataFrame(record)
```

Out[7]:

	DbList
0	pubmed
1	protein
2	nuccore
3	ipg
4	nucleotide
5	structure
6	genome
7	annotinfo
8	assembly
9	bioproject
10	biosample
11	blastdbinfo
12	books
13	cdd
14	clinvar
15	gap
16	gapplus
17	grasp
18	dbvar
19	gene
20	gds
21	geoprofiles
22	homologene
23	medgen
24	mesh
25	ncbisearch
26	nlmcatalog
27	omim
28	orgtrack
29	pmc
30	popset
31	proteinclusters
32	pcassay
33	protfam

	DbList
34	biosystems
35	pccompound
36	pcsubstance
37	seqannot
38	snp
39	sra

Για να δούμε περισσότερες πληροφορίες για όλες αυτές τις βάσεις μπορούμε να πάμε στο λινκ:

https://www.ncbi.nlm.nih.gov/{ONOMA_ΒΑΣΗΣ} (https://www.ncbi.nlm.nih.gov/{ONOMA_ΒΑΣΗΣ})

Π.χ:

- https://www.ncbi.nlm.nih.gov/sra (https://www.ncbi.nlm.nih.gov/sra)
- https://www.ncbi.nlm.nih.gov/probe (https://www.ncbi.nlm.nih.gov/probe)
- κτλ...

Κάθε στοιχείο της βάσης αυτή έχει ένα "id". Μπορούμε να "κατεβάσουμε" μία εγγραφή της βάσης δίνοντας το ονομα της βάσης, το id του στοιχείου και σε τι φορμάτ το θέλουμε. Για παράδειγμα μπορούμε να κατεβάσουμε σε fasta φορμάτ την ακολουθία με id: "NM_000762.5" από τη nuccore βάση. Το link αυτής της εγγραφής είναι το εξής: https://www.ncbi.nlm.nih.gov/nuccore/189339232 (https://www.ncbi.nlm.nih.gov/nuccore/189339232)

In [8]:
```
handle = Entrez.efetch(
  db='nuccore',
  id='NM_000762.5',
  retmode='text',
  rettype='fasta',
  #rettype='genbank',
)

data = handle.read()
print (data)
```

```
>NM_000762.5 Homo sapiens cytochrome P450 family 2 subfamily A mem
ber 6 (CYP2A6), mRNA
ATCTATCATCCCACTACCACCATGCTGGCCTCAGGGATGCTTCTGGTGGCCTTGCTGGTCTGCCTG
ACTG
TAATGGTCTTGATGTCTGTTTGGCAGCAGAGGAAGAGCAAGGGGAAGCTGCCTCCGGGACCCACCC
CATT
GCCCTTCATTGGAAACTACCTGCAGCTGAACACAGAGCAGATGTACAACTCCCTCATGAAGATCAG
TGAG
CGCTATGGCCCCGTGTTCACCATTCACTTGGGGCCCCGGCGGGTCGTGGTGCTGTGTGGACATGAT
GCCG
TCAGGGAGGCTCTGGTGGACCAGGCTGAGGAGTTCAGCGGGCGAGGCGAGCAAGCCACCTTCGACT
GGGT
CTTCAAAGGCTATGGCGTGGTATTCAGCAACGGGGAGCGCGCCAAGCAGCTCCGGCGCTTCTCCAT
CGCC
ACCCTGCGGGACTTCGGGGTGGGCAAGCGAGGCATCGAGGAGCGCATCCAGGAGGAGGCGGGCTTC
CTCA
TCGACGCCCTCCGGGGCACTGGCGGCGCCAATATCGATCCCACCTTCTTCCTGAGCCGCACAGTCT
CCAA
TGTCATCAGCTCCATTGTCTTTGGGGACCGCTTTGACTATAAGGACAAAGAGTTCCTGTCACTGTT
GCGC
ATGATGCTAGGAATCTTCCAGTTCACGTCAACCTCCACGGGGCAGCTCTATGAGATGTTCTCTTCG
GTGA
TGAAACACCTGCCAGGACCACAGCAACAGGCCTTTCAGTTGCTGCAAGGGCTGGAGGACTTCATAG
CCAA
GAAGGTGGAGCACAACCAGCGCACGCTGGATCCCAATTCCCCACGGGACTTCATTGACTCCTTTCT
CATC
CGCATGCAGGAGGAGGAGAAGAACCCCAACACGGAGTTCTACTTGAAAAACCTGGTGATGACCACG
TTGA
ACCTCTTCATTGGGGGCACCGAGACCGTCAGCACCACCCTGCGCTATGGCTTCTTGCTGCTCATGA
AGCA
CCCAGAGGTGGAGGCCAAGGTCCATGAGGAGATTGACAGAGTGATCGGCAAGAACCGGCAGCCCAA
GTTT
GAGGACCGGGCCAAGATGCCCTACATGGAGGCAGTGATCCACGAGATCCAAAGATTTGGAGACGTG
ATCC
CCATGAGTTTGGCCCGCAGAGTCAAAAAGGACACCAAGTTTCGGGATTTCTTCCTCCCTAAGGGCA
CCGA
AGTGTACCCTATGCTGGGCTCTGTGCTGAGAGACCCCAGTTTCTTCTCCAACCCCCAGGACTTCAA
TCCC
CAGCACTTCCTGAATGAGAAGGGGCAGTTTAAGAAGAGTGATGCTTTTGTGCCCTTTTCCATCGGA
AAGC
GGAACTGTTTCGGAGAAGGCCTGGCCAGAATGGAGCTCTTTCTCTTCTTCACCACCGTCATGCAGA
ACTT
CCGCCTCAAGTCCTCCCAGTCACCTAAGGACATTGACGTGTCCCCCAAACACGTGGGCTTTGCCAC
GATC
CCACGAAACTACACCATGAGCTTCCTGCCCCGCTGAGCGAGGGCTGTGCCGGTGCAGGTCTGGTGG
GCGG
GGCCAGGGAAAGGGCAGGGCCAAGACCGGGCTTGGGAGAGGGGCGCAGCTAAGACTGGGGGCAGGA
TGGC
GGAAAGGAAGGGGCGTGGTGGCTAGAGGGAAGAGAAGAAACAGAAGCGGCTCAGTTCACCTTGATA
AGGT
GCTTCCGAGCTGGGATGAGAGGAAGGAAACCCTTACATTATGCTATGAAGAGTAGTAATAATAGCA
GCTC
TTATTTCCTGAGCAAAAAAAAAAAA
```

Μπορούμε τώρα να τρέξουμε BLAT (https://en.wikipedia.org/wiki/BLAT_(bioinformatics%29) σε αυτή την ακολουθία και να δούμε σε ποιο σημείο του ανθρώπινου γονιδιώματος ανοίκει:

In [9]:
```
from Bio.Blast import NCBIWWW
from Bio.Blast import NCBIXML

#result_handle = NCBIWWW.qblast("blastn", "refseq_genomic_human", d
ata)
result_handle = NCBIWWW.qblast("blastn", "nt", data.format("fast
a"))
blast_record = NCBIXML.read(result_handle)
```

Σε πόσες ακολουθίες έγινε match;

In [10]:
```
len(blast_record.alignments)
```

Out[10]:
```
50
```

Ποιες είναι αυτές;

In [11]:
```
for x in blast_record.alignments:
    print (x.title)
```

```
gi|1519246407|ref|NM_000762.6| Homo sapiens cytochrome P450 family
2 subfamily A member 6 (CYP2A6), mRNA
gi|180986|gb|M33318.1|HUMCPIIA3A Human cytochrome P450IIA3 (CYP2A
3) mRNA, complete cds
gi|29546|emb|X13897.1| Human mRNA for cytochrome P-450IIA
gi|30331|emb|X13930.1| Human CYP2A4 mRNA for P-450 IIA4 protein
gi|1753053698|ref|XM_019015583.2| PREDICTED: Gorilla gorilla goril
la cytochrome P450 2A6 (LOC101146638), transcript variant X1, mRNA
gi|64654819|gb|BC096256.1| Homo sapiens cytochrome P450, family 2,
subfamily A, polypeptide 6, mRNA (cDNA clone MGC:116921 IMAGE:4000
6068), complete cds
gi|64654814|gb|BC096255.1| Homo sapiens cytochrome P450, family 2,
subfamily A, polypeptide 6, mRNA (cDNA clone MGC:116920 IMAGE:4000
6064), complete cds
gi|64653226|gb|BC096254.1| Homo sapiens cytochrome P450, family 2,
subfamily A, polypeptide 6, mRNA (cDNA clone MGC:116919 IMAGE:4000
6062), complete cds
gi|109730085|gb|BC096253.3| Homo sapiens cytochrome P450, family
2, subfamily A, polypeptide 6, mRNA (cDNA clone MGC:116918 IMAGE:4
0006061), complete cds
gi|1351473687|ref|XM_024237332.1| PREDICTED: Pongo abelii cytochro
me P450 2A6 (LOC100457048), transcript variant X1, mRNA
gi|35197|emb|X13929.1| Human CYP2A3 mRNA for P-450 IIA3 protein
gi|6470138|gb|AF182275.1|AF182275 Homo sapiens cytochrome P450-2A6
(CYP2A6) mRNA, complete cds
gi|1887789759|ref|NM_000764.3| Homo sapiens cytochrome P450 family
2 subfamily A member 7 (CYP2A7), transcript variant 1, mRNA
gi|1849054764|ref|XM_034944594.1| PREDICTED: Pan paniscus cytochro
me P450 family 2 subfamily A member 7 (CYP2A7), mRNA
gi|181269|gb|M33317.1|HUMCYIIA4A Human cytochrome P450IIA4 (CYP2A
4) mRNA, complete cds
gi|1367220708|ref|XM_024351394.1| PREDICTED: Pan troglodytes cytoc
hrome P450 2A7 (LOC107966456), mRNA
gi|1008465|gb|U22029.1|HSU22029 Human cytochrome P450 (CYP2A7) mRN
A, complete cds
gi|1743170194|ref|XM_003282448.4| PREDICTED: Nomascus leucogenys c
ytochrome P450 2A13 (LOC100579248), transcript variant X1, mRNA
gi|1800015647|ref|XM_032172847.1| PREDICTED: Hylobates moloch cyto
chrome P450 2A13 (LOC116480658), transcript variant X1, mRNA
gi|1751200936|ref|XM_010380822.2| PREDICTED: Rhinopithecus roxella
na cytochrome P450 2A13-like (LOC104676097), transcript variant X
1, mRNA
gi|795398387|ref|XM_012086802.1| PREDICTED: Cercocebus atys cytoch
rome P450 2A13-like (LOC105598002), transcript variant X1, mRNA
gi|1825863830|ref|XM_033224912.1| PREDICTED: Trachypithecus franco
isi cytochrome P450 2A13 (LOC117091930), mRNA
gi|548960720|ref|NM_001285348.1| Macaca fascicularis cytochrome P4
50 family 2 subfamily A member 24 (CYP2A24), mRNA >gi|71152698|gb|
DQ074792.1| Macaca fascicularis cytochrome P450 2A24 (CYP2A24) mRN
A, complete cds
gi|164691768|dbj|AK312964.1| Homo sapiens cDNA, FLJ93424, highly s
imilar to Homo sapiens cytochrome P450, family 2, subfamily A, pol
ypeptide 6 (CYP2A6), mRNA
gi|1825821225|ref|XM_033218604.1| PREDICTED: Trachypithecus franco
isi cytochrome P450 2A13-like (LOC117087862), mRNA
gi|1653961540|ref|NM_000766.5| Homo sapiens cytochrome P450 family
2 subfamily A member 13 (CYP2A13), mRNA
gi|1411126738|ref|XM_025366233.1| PREDICTED: Theropithecus gelada
cytochrome P450 2A13-like (LOC112612112), transcript variant X2, m
```

```
RNA
gi|795398405|ref|XM_012086807.1| PREDICTED: Cercocebus atys cytoch
rome P450 2A13-like (LOC105598006), mRNA
gi|1351473349|ref|XM_024237225.1| PREDICTED: Pongo abelii cytochro
me P450 2A13 (LOC100458875), transcript variant X1, mRNA
gi|795234893|ref|XM_011952499.1| PREDICTED: Colobus angolensis pal
liatus cytochrome P450 2A13-like (LOC105519195), transcript varian
t X1, mRNA
gi|1622890953|ref|XM_028838334.1| PREDICTED: Macaca mulatta cytoch
rome P450, family 2, subfamily A, polypeptide 24 (CYP2A24), transc
ript variant X1, mRNA
gi|1849062391|ref|XM_003812465.2| PREDICTED: Pan paniscus cytochro
me P450 family 2 subfamily A member 13 (CYP2A13), mRNA
gi|1788688032|ref|XM_023229274.3| PREDICTED: Piliocolobus tephrosc
eles cytochrome P450 2A13 (LOC111554025), transcript variant X1, m
RNA
gi|1411126736|ref|XM_025366232.1| PREDICTED: Theropithecus gelada
cytochrome P450 2A13-like (LOC112612112), transcript variant X1, m
RNA
gi|649118244|gb|KJ896677.1| Synthetic construct Homo sapiens clone
ccsbBroadEn_06071 CYP2A6 gene, encodes complete protein
gi|1147694741|emb|LT740833.1| Human ORFeome Gateway entry vector p
ENTR223-CYP2A6, complete sequence
gi|823673899|gb|KR711779.1| Synthetic construct Homo sapiens clone
CCSBHm_00030839 CYP2A6 (CYP2A6) mRNA, encodes complete protein
gi|823673897|gb|KR711778.1| Synthetic construct Homo sapiens clone
CCSBHm_00030834 CYP2A6 (CYP2A6) mRNA, encodes complete protein
gi|823673895|gb|KR711777.1| Synthetic construct Homo sapiens clone
CCSBHm_00030831 CYP2A6 (CYP2A6) mRNA, encodes complete protein
gi|823673893|gb|KR711776.1| Synthetic construct Homo sapiens clone
CCSBHm_00030829 CYP2A6 (CYP2A6) mRNA, encodes complete protein
gi|795398351|ref|XM_012086792.1| PREDICTED: Cercocebus atys cytoch
rome P450 2A13 (LOC105597997), transcript variant X3, mRNA
gi|795398348|ref|XM_012086791.1| PREDICTED: Cercocebus atys cytoch
rome P450 2A13 (LOC105597997), transcript variant X2, mRNA
gi|795398345|ref|XM_012086790.1| PREDICTED: Cercocebus atys cytoch
rome P450 2A13 (LOC105597997), transcript variant X1, mRNA
gi|544525941|ref|XM_005595716.1| PREDICTED: Macaca fascicularis cy
tochrome P450 2A13-like (LOC102143277), mRNA
gi|1751208157|ref|XM_030913665.1| PREDICTED: Rhinopithecus roxella
na cytochrome P450 2A13 (LOC104676180), mRNA
gi|1753052882|ref|XM_031004339.1| PREDICTED: Gorilla gorilla goril
la cytochrome P450 2A13 (LOC101127384), mRNA
gi|795137501|ref|XM_011936512.1| PREDICTED: Colobus angolensis pal
liatus cytochrome P450 2A13 (LOC105507828), transcript variant X1,
mRNA
gi|1825820915|ref|XM_033216130.1| PREDICTED: Trachypithecus franco
isi cytochrome P450 2A13-like (LOC117086673), mRNA
gi|1751201107|ref|XM_010380820.2| PREDICTED: Rhinopithecus roxella
na cytochrome P450 2A13 (LOC104676095), mRNA
gi|1381460778|ref|XM_011764694.2| PREDICTED: Macaca nemestrina cyt
ochrome P450 2A13 (LOC105495313), mRNA
```

Ας πάρουμε το πρώτο:

In [12]:
```
first_alignment = blast_record.alignments[0]
```

Το κάθε alignment έχει πολλά "hsps". Σύμφωνα με το documentation (http://biopython.org/DIST/docs/tutorial/Tutorial.html#htoc105):

> HSP (high-scoring pair) represents region(s) in the hit sequence that contains significant alignment(s) to the query sequence. It contains the actual match between your query sequence and a database entry. As this match is determined by the sequence search tool's algorithms, the HSP object contains the bulk of the statistics computed by the search tool. This also makes the distinction between HSP objects from different search tools more apparent compared to the differences you've seen in Query Result or Hit objects.

Ας πάρουμε το πρώτο:

In [13]:
```
first_hsp = first_alignment.hsps[0]
```

Μπορούμε να τυπώσουμε το match:

In [14]:
```
print (first_hsp)
```

```
Score 3522 (3177 bits), expectation 0.0e+00, alignment length 1761
Query:       1 ATCTATCATCCCACTACCACCATGCTGGCCTCAGGGATGCTTCTG...TGA
1761
                 ||||||||||||||||||||||||||||||||||||||||||...|||
Sbjct:       1 ATCTATCATCCCACTACCACCATGCTGGCCTCAGGGATGCTTCTG...TGA
1761
```

Επίσης μπορούμε να πάρουμε πληροφορίες για το match:

In [15]:
```
print ('Subject start:', first_hsp.sbjct_start)
print ('Subject end:', first_hsp.sbjct_end)
print ('Query start:', first_hsp.query_start)
print ('Query end:', first_hsp.query_end)
```

```
Subject start: 1
Subject end: 1761
Query start: 1
Query end: 1761
```

Ας δούμε πως είναι και ένα "κακό" alignment:

In [16]:
```
last_alignment = blast_record.alignments[-1]
last_hsp = last_alignment.hsps[-1]
print (last_hsp)
```

```
Score 2873 (2591 bits), expectation 0.0e+00, alignment length 1765
Query:       1 ATCTATCATCCCACTACCACCATGCTGGCCTCAGGGATGCTTCTG...CAA
1765
                 ||| |||||||||||  || |||||||||||||||| |||| |||...|||
Sbjct:     564 ATCGATCATCCCACTGCCCCCATGCTGGCCTCAGGGCTGCTCCTG...CAA
2310
```

Paiwise alignment με biopython

In [17]:
```
from Bio import pairwise2
from Bio.pairwise2 import format_alignment

alignments = pairwise2.align.globalxx("ACCGT", "ACG")
print(format_alignment(*alignments[0]))
```

```
ACCGT
| ||
A-CG-
  Score=3
```

Άλλες βάσεις δεδομένων

Γενικότερα το τοπίο με της online βάσεις γενομικών δεδομένων είναι αρκετά δυναμικό. Καινούργιες βάσεις προστίθονται, νέα APIs κτλ. Σε επίπεδα οργανισμών υπάρχουν δύο μεγάλοι το NCBI και το EMBL. Το NCBI έχει ομαδοποιήσει όλες τις βάσεις δεδομένων με το σύστημα Entrez, ενώ το EMBL έχει το REST API (https://rest.ensembl.org/). Άλλες αντιστοιχίες είναι:

Service	EMBL (https://www.embl.de/)	NCBI (https://www.ncbi.nlm.nih.gov/)
Variant Annotation	VEP (https://www.ensembl.org/info/docs/tools/vep/index.html)	Variation Reporter (https://www.ncbi.nlm.nih.gov/variation/tools/reporter)
Raw Sequencing Data	European Nucleotide Archive (https://www.ebi.ac.uk/ena)	SRA (https://www.ncbi.nlm.nih.gov/sra)
Expression Data	ArrayExpress (https://www.ebi.ac.uk/arrayexpress/)	Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/)
Phenotype	EGA (European Genome-Phenome Archive) (https://www.ebi.ac.uk/ega/home)	dbGAP (https://www.ncbi.nlm.nih.gov/gap)
Mutation Database	Ensembl (http://www.ensembl.org/index.html)	dbSNP (https://www.ncbi.nlm.nih.gov/projects/SNP/)
Genomic Browsers	Ensembl (http://www.ensembl.org/Homo_sapiens/Location/View?r=17:63992802-64038237)	UCSC (https://genome.ucsc.edu/cgi-bin/hgTracks?db=hg38&lastVirtModeType=default&lastVirtModeExtraState=&virtModeType=default&virtMode=0&nonVirtPosition=&position=chr17%3A63992802%2D64038237&hgsid=654038027_kzafCuBRWkvSS4Uq55z0pjldJcyS)
Location converters	REST API (map) (https://rest.ensembl.org/documentation/info/assembly_map)	remap (https://www.ncbi.nlm.nih.gov/genome/tools/remap)
Query services	REST API (https://rest.ensembl.org/)	Entrez (https://www.ncbi.nlm.nih.gov/Class/MLACourse/Original8Hour/Entrez/)

Εκτός από ENSEMBL/NCBI μια άλλη πολύ καλή βάση είναι:

- MyGene (http://mygene.info/), MyVariant (http://myvariant.info/) . Πολύ μοντέρνο API, φιλικό και καλή οργάνωση πληροφορίας. Δοκιμάστε το: http://myvariant.info/v1/api (http://myvariant.info/v1/api) . Το αρνητικό του είναι ότι περιέχει ΜΟΝΟ πληροφορίες για variants (π.χ. δεν έχει transcripts, proteins, ...)