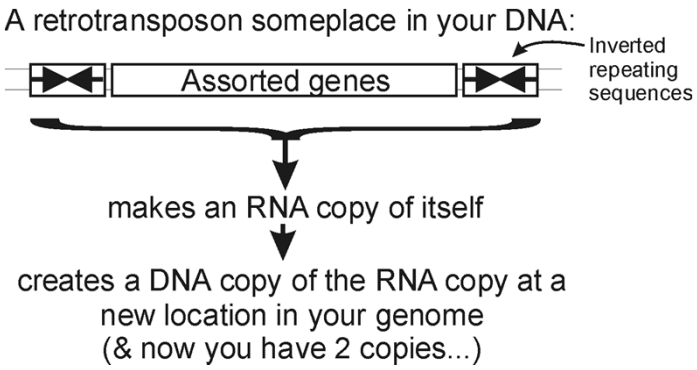
You and your (DNA) parasites



Events like these, happening over and over again, have led to…

You and your (DNA) parasites

Major types of repeats in the human genome

			Length	Copies	Fraction of genome
LINEs	Autonomous	ORF1 ORF2 (pol) AAA	6-8 kb	850,000	21%
SINEs	Non-autonomous	A B AAA	100-300 bp	1,500,000	13%
Retrovirus-like elements	Autonomous	gag pol (env)	6-11 kb	450,000	8%
	Non-autonomous	(gag)	1.5-3 kb		
DNA transposon fossils	Autonomous	transposase	2-3 kb	300,000	3%
	Non-autonomous		80-3,000 bp		
					~45%

Bottom line: Roughly half of your (and my) genome is the fossil wreckage of genomic parasites.

We know this (in part) from sequence alignments.

So far, we've talked about

- DNA, RNA and protein sequences
- How to compare sequences to decide if they are related
- Having databases full of sequences and comparing them rapidly (BLAST)

In fact, <u>many</u> such databases exist, so today we'll start with a brief tour of <u>some</u> of the biological data on the web.

Just some of the resources available for bioinformatics

Think of these as the raw data for new discoveries

Database	Records	Address
BioGRID	1.6 M protein interactions	https://thebiogrid.org
EcoCyc/MetaCyc	>2,600 pathways	http://www.ecocyc.org, http://www.metacyc.org
Ensembl	Major repository of DNA sequences, genomes, genes, proteins, and transcripts	http://useast.ensembl.org/index.html
Entrez Genome	Thousands of genome sequences	http://www.ncbi.nlm.nih.gov/genome?db=genome
Genbank	>285 billion bases sequenced; > 3 trillion bases as whole genome shotgun data	https://www.ncbi.nlm.nih.gov/genbank/
Gene Expression Omnibus (GEO)	>2.8M mRNA or protein expression data sets	http://www.ncbi.nlm.nih.gov/geo/
Genomes Online Database (GOLD)	>150K genome sequences, many in progress	https://gold.jgi.doe.gov/index
Human Protein Atlas	millions of high-res images of ~17K human proteins across tissues, cancers, & cell lines	http://www.proteinatlas.org/
KEGG	Most known pathways, in 530 graphical diagrams and >6K organisms (via homology)	http://www.genome.ad.jp/kegg/
Medline	>29 million references	https://www.ncbi.nlm.nih.gov/PubMed/
Mouse Genome Informatics	~20,000 mouse genes, diverse associated data & annotations	http://www.informatics.jax.org/
Online Mendelian Inheritance in Man (OMIM)	Compendium of human genes and genetic phenotypes, data for >16,000 human genes	https://www.ncbi.nlm.nih.gov/omim/
Pride	> 342 million peptide mass spectra from 27K experiments	https://www.ebi.ac.uk/pride/archive/
Reactome	>2K pathways involving >10K human proteins, also other organisms	https://www.reactome.org/
SGD	~6,000 yeast genes, diverse associated data & annotations	https://www.yeastgenome.org/
UniProtKB/SWISS-PROT	>500K hand-curated sequence entries from >9K organisms	https://www.uniprot.org/

Just some of the resources available for bioinformatics

Think of these as the raw data for new discoveries

Database	Record	
BioGRID	1.6 M protein interactions	
EcoCyc/MetaCyc	>2,600 pathways	
Ensembl	Major repository of DNA sequences, genomes, genes, proteins, and transcripts	http://useast.ensembl.org/index.html
Entrez Genome	Thousands of genome sequences	
Genbank	>285 billion bases sequenced; > 3 trillion bases as whole genome shotgun data	
Gene Expression Omnibus (GEO)	>2.8M mRNA or protein expression data sets	
Genomes Online Database (GOLD)	>150K genome sequences, many in progress	
Human Protein Atlas		http://www.proteinatlas.org/
KEGG		
Medline		
Mouse Genome Informatics	annotations	
Online Mendelian Inheritance in Man (OMIM)	Compendium of human genes and genetic phenotypes, data for >16,000 human genes	
Pride	> 342 million peptide mass spectra from 27K experiments	https://www.ebi.ac.uk/pride/archive/
Reactome	>2K pathways involving >10K human protein also other organisms	
SGD	~6,000 yeast genes, diverse associated data annotations	
UniProtKB/SWISS-PROT	>500K hand-curated sequence entries from >9K organisms	https://www.uniprot.org/

Biogrid has 1.6 M protein-protein interactions (https://thebiogrid.org/)

GEO has millions of experiments, each measuring 1000's of mRNA or protein abundances

Medline has >29 million research articles, many with complete text online

OMIM = the most important resource for human genetic disease

Uniprot = a frequent first step to learn about genes

Live demo OMIM,
Reactome,
Human Protein Atlas

It's nice to know that all of this exists, but ideally, you'd like to be able to so something constructive with the data.

That means getting the data inside your own programs.

All of these databases let you download data in big batches, but this isn't always the case, so….

Let's empower your Python scripts to grab data from the web.

We'll use Python library/module = an optional, specialized set of Python methods

This particular Python module is called ***urllib2***.

urllib2 is:

- A collection of programs/tools to let you to surf the web from inside your programs.
- Much more powerful than the simple tasks we'll do with it.
- More details: http://docs.python.org/2/library/urllib2.html

The basic idea:

We first set up a “request” by opening a connection to the URL.

We then save the response in a variable and print it.

If it can’t connect to the site, it’ll print out a helpful error message instead of the page.

You can more or less use the commands in a cookbook fashion….

For example:

```
import urllib2                          # include the urllib2 module

url = "http://www.marcottelab.org/index.php/BCH394P_BCH364C_2019"

try:                    # this 'try' statement tells Python that we might expect an error.
    request = urllib2.urlopen(url)      # setup a request
    page = request.read()               # save the response
    print page                          # show the result to the user

except urllib2.HTTPError:               # handle a page not found error
    print "Could not find page."
```

→ **Run this…**

→ **We just captured the class web page and printed it out…**

```
>>>
<!DOCTYPE html>
<html lang="en" dir="ltr" class="client-nojs">
<head>
<title>BCH394P BCH364C 2019 - Marcotte Lab</title>
<meta charset="UTF-8" />
<meta name="generator" content="MediaWiki 1.21.2" />
<link rel="shortcut icon" href="/favicon.ico" />
<link rel="search" type="application/opensearchdescription+xml"
href="/opensearch_desc.php" title="Marcotte Lab (en)" />
<link rel="EditURI" type="application/rsd+xml"
href="http://www.marcottelab.org/api.php?action=rsd" />
<link rel="copyright" href="http://creativecommons.org/licenses/by-nc-nd/3.0/" />
<link rel="alternate" type="application/atom+xml" title="Marcotte Lab Atom feed“
```

…and so on, and on, and on…

That was a static web page.

Let’s try one that requires some sort of action, for example by entering a document id or an id code for a sequence.

Many web pages pass this information along in the web URL itself…

Here's a complete Python program to retrieve a single entry from Medline:

```
import urllib2
pmid = 11237011

# Insert the pmid where the {} are in the following URL:
url = "https://www.ncbi.nlm.nih.gov/pubmed/{0}?report=medline&format=text".format(pmid)

try:                                # there might be an error!
    request = urllib2.urlopen(url)
    page = request.read()
    print page

except urllib2.HTTPError:       # handle page not found error
    print "Could not connect to Medline!"
```

If you run that program, you should get back…

```
>>>
<?xml version="1.0" encoding="utf-8"?>
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN" "http://www.w3.org/TR/xhtml1/DTD/xhtml1-
transitional.dtd">
<pre>
PMID- 11237011
OWN - NLM
STAT- MEDLINE
DA  - 20010309
DCOM- 20010322
LR  - 20061115
IS  - 0028-0836 (Print)
IS  - 0028-0836 (Linking)
VI  - 409
IP  - 6822
DP  - 2001 Feb 15
TI  - Initial sequencing and analysis of the human genome.
PG  - 860-921
AB  - The human genome holds an extraordinary trove of information about human
      development, physiology, medicine and evolution. Here we report the results of an
      international collaboration to produce and make freely available a draft sequence
      of the human genome. We also present an initial analysis of the data, describing
      some of the insights that can be gleaned from the sequence.
FAU - Lander, E S
AU  - Lander ES
AD  - Whitehead Institute for Biomedical Research, Center for Genome Research,
      Cambridge, Massachusetts 02142, USA. lander@genome.wi.mit.edu
```

[and so on]

the Medline entry for the human genome sequence paper

If you run that program, you should get back…

```
>>>
<?xml version="1.0" encoding="utf-8"?>
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN" "http://www.w3.org/TR/xhtml1/DTD/xhtml1-
transitional.dtd">
<pre>
PMID- 11237011
OWN - NLM
STAT- MEDLINE
DA  - 20010309
DCOM- 20010322
LR  - 20061115
IS  - 0028-0836 (Print)
IS  - 0028-0836 (Linking)
VI  - 409
IP  - 6822
DP  - 2001 Feb 15
TI  - Initial sequencing and analysis of the human genome.
PG  - 860-921
AB  - The human genome holds an extraordinary trove of information about human
      development, physiology, medicine and evolution. Here we report the results of an
      international collaboration to produce and make freely available a draft sequence
      of the human genome. We also present an initial analysis of the data, describing
      some of the insights that can be gleaned from the sequence.
FAU - Lander, E S
AU  - Lander ES
AD  - Whitehead Institute for Biomedical Research, Center for Genome Research,
      Cambridge, Massachusetts 02142, USA. lander@genome.wi.mit.edu
```

[and so on]

We just printed it. We could have saved it or extracted data from it. For example…

Here's our Python program again to retrieve a single entry from Medline. How would we modify this to count the authors?

```
import urllib2
pmid = 11237011

# Insert the pmid where the {} are in the following URL:
url = "https://www.ncbi.nlm.nih.gov/pubmed/{0}?report=medline&format=text".format(pmid)

try:                                # there might be an error!
    request = urllib2.urlopen(url)
    page = request.read()
    print page

except urllib2.HTTPError:       # handle page not found error
    print "Could not connect to Medline!"
```

Here's our Python program again to retrieve a single entry from Medline. How would we modify this to count the authors?

```
import urllib2
pmid = 11237011

# Insert the pmid where the {} are in the following URL:
url = "https://www.ncbi.nlm.nih.gov/pubmed/{0}?report=medline&format=text".format(pmid)

try:                                # there might be an error!
    request = urllib2.urlopen(url)
    page = request.read()
    print page.count("AU  - ")

except urllib2.HTTPError:       # handle page not found error
    print "Could not connect to Medline!"
```

Medline begins author lines with "AU - " , so…

→ **Run this, & get …**

```
>>>
256
```

So, there were 256 authors on one (of the two) human genome papers

- Queries to Medline or any other NCBI database, including GenBank, are described at:
  http://www.ncbi.nlm.nih.gov/books/NBK3862/
  (& for that matter, all of medline is downloadable)
- You can often figure out the form of the URL just by looking something up in a database, then noting the address of the web page with the data.
- This very simple approach could easily be the basis for:
  - a home-made web browser
  - a program to consult biological databases in real time
  - a program to map the internet, etc.
- Of course, with this kind of power available to you, the imagination reels...