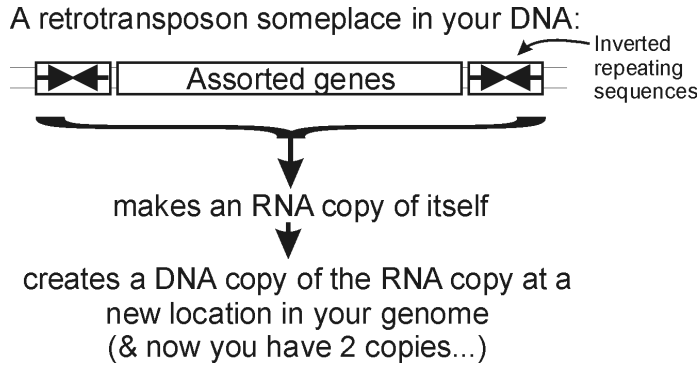


# You and your (DNA) parasites



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

1

# You and your (DNA) parasites

Major types of repeats in the human genome

			Length	Copies	Fraction of genome
<b>LINES</b>	<b>Autonomous</b>		6-8 kb	850,000	21%
<b>SINES</b>	<b>Non-autonomous</b>		100-300 bp	1,500,000	13%
<b>Retrovirus-like elements</b>	<b>Autonomous</b>		6-11 kb	450,000	8%
	<b>Non-autonomous</b>		1.5-3 kb		
<b>DNA transposon fossils</b>	<b>Autonomous</b>		2-3 kb	300,000	3%
	<b>Non-autonomous</b>		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.**

2

So far, we've talked about

- DNA, RNA and protein sequences
  - hang on! We didn't talk about RNA!
  - Can we align RNA the same way as DNA & proteins?
- How to compare sequences to decide if they are related
- Having databases full of sequences and comparing them rapidly (BLAST)

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

3

	Database	Records	Address
<p>Just some of the resources available for bioinformatics</p> <p>Think of these as the raw data for new discoveries</p>	BioGRID	>2 M protein interactions	<a href="https://thebiogrid.org">https://thebiogrid.org</a>
	EcoCyc/MetaCyc	>2,700 pathways from >3,000 organisms	<a href="http://www.ecocyc.org">http://www.ecocyc.org</a> , <a href="http://www.metacyc.org">http://www.metacyc.org</a>
	Ensembl (+ BioMart for easy sequence queries)	Major repository of DNA sequences, genomes, genes, proteins, and transcripts	<a href="http://useast.ensembl.org/index.html">http://useast.ensembl.org/index.html</a>
	Entrez Genome	Thousands of genome sequences	<a href="http://www.ncbi.nlm.nih.gov/genome?db=genome">http://www.ncbi.nlm.nih.gov/genome?db=genome</a>
	Expression Atlas	139K mRNA expression expts in 65 species	<a href="https://www.ebi.ac.uk/gxa/home/">https://www.ebi.ac.uk/gxa/home/</a>
	Genbank	>1 trillion bases sequenced; > 14 trillion bases as whole genome shotgun data	<a href="https://www.ncbi.nlm.nih.gov/genbank/">https://www.ncbi.nlm.nih.gov/genbank/</a>
	Gene Expression Omnibus (GEO)	>4 M mRNA or protein expression expts	<a href="http://www.ncbi.nlm.nih.gov/geo/">http://www.ncbi.nlm.nih.gov/geo/</a>
	Genomes Online Database (GOLD)	>180K genome sequences, many in progress	<a href="https://gold.jgi.doe.gov/index">https://gold.jgi.doe.gov/index</a>
	Human Protein Atlas	millions of high-res images of ~17K human proteins across tissues, cancers, & cell lines	<a href="http://www.proteinatlas.org/">http://www.proteinatlas.org/</a>
	KEGG	Most known pathways, in 548 graphical diagrams and >7K organisms (via homology)	<a href="http://www.genome.ad.jp/kegg/">http://www.genome.ad.jp/kegg/</a>
	Medline / PubMed	>30 million references	<a href="https://www.ncbi.nlm.nih.gov/pubmed/">https://www.ncbi.nlm.nih.gov/pubmed/</a>
	Mouse Genome Informatics	~20,000 mouse genes, diverse associated data & annotations	<a href="http://www.informatics.jax.org/">http://www.informatics.jax.org/</a>
	Online Mendelian Inheritance in Man (OMIM)	Compendium of human genes and genetic phenotypes, data for >16,000 human genes	<a href="https://www.ncbi.nlm.nih.gov/omim/">https://www.ncbi.nlm.nih.gov/omim/</a>
	Pride	Hundreds of millions of peptide mass spectra from 10's of thousands of experiments	<a href="https://www.ebi.ac.uk/pride/archive/">https://www.ebi.ac.uk/pride/archive/</a>
	Reactome	>2K pathways involving >10K human proteins, also other organisms	<a href="https://www.reactome.org/">https://www.reactome.org/</a>
	SGD	~6,000 yeast genes, diverse associated data & annotations	<a href="https://www.yeastgenome.org/">https://www.yeastgenome.org/</a>
	UniProtKB/SWISS-PROT	>550K hand-curated sequence entries from >14K organisms	<a href="https://www.uniprot.org/">https://www.uniprot.org/</a>

4

**Just some of the resources available for bioinformatics**

**Think of these as the raw data for new discoveries**

Database	Records	URL
BioGRID	>2 M protein interactions	
EcoCyc/MetaCyc	>2,700 pathways from >3,000 organisms	
Ensembl (+ BioMart for easy sequence queries)	Major repository of DNA sequences, genes, proteins, and transcripts	
Entrez Genome	Thousands of genome sequences	<a href="http://www.ncbi.nlm.nih.gov/genome?db=genome">http://www.ncbi.nlm.nih.gov/genome?db=genome</a>
Expression Atlas	121K mRNA expression expts in 62 species	
Genbank	>386 billion bases sequenced; > 5 trillion bases as whole genome shotgun data	
Gene Expression Omnibus (GEO)	>4 M mRNA or protein expression expts	
Genomes Online Database (GOLD)	>150K genome sequences, many in progress	
Human Protein Atlas		<a href="http://www.proteinatlas.org/">http://www.proteinatlas.org/</a>
KEGG		<a href="http://www.genome.jp/kegg/">http://www.genome.jp/kegg/</a>
Medline / PubMed		<a href="http://pubmed.ncbi.nlm.nih.gov/">http://pubmed.ncbi.nlm.nih.gov/</a>
Mouse Genome Informatics	~20,000 mouse genes, diverse associated data & annotations	<a href="http://www.informatics.jax.org/">http://www.informatics.jax.org/</a>
Online Mendelian Inheritance in Man (OMIM)	Compendium of human genes and genetic phenotypes, data for >16,000 human genes	<a href="http://www.omim.org/">http://www.omim.org/</a>
Pride	Hundreds of protein mass spectrometry datasets from 10's of organisms	<a href="http://www.ebi.ac.uk/pride/">http://www.ebi.ac.uk/pride/</a>
Reactome	>2K pathways, also other biological processes	<a href="http://www.reactome.org/">http://www.reactome.org/</a>
SGD	~6,000 yeast genes, annotated with protein interactions	<a href="http://www.yeastgenome.org/">http://www.yeastgenome.org/</a>
UniProtKB/SWISS-PROT	>550K high quality protein sequences from all organisms	<a href="http://www.uniprot.org/">http://www.uniprot.org/</a>

**Biogrid has >2 M protein-protein interactions (<https://thebiogrid.org/>)**

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

**Medline has >30 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. Also **amazingly useful** for interconverting IDs and linking to other resources**

5

Live demo Ensembl->BioMart->filter for [IPR031588], OMIM, Reactome, Human Protein Atlas

6

**It's nice to know that all of this exists, but ideally, you'd like to be able to do 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....**

7

**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 ***urllib*** (Py3) or ***urllib2*** (Py2)

urllib/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:

<https://docs.python.org/3.8/library/urllib.request.html>

or <http://docs.python.org/2/library/urllib2.html>

8

### 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....

9

### For example:

```
import urllib.request          # include the urllib.request module

url = "https://www.utexas.edu/"

x = urllib.request.urlopen(url) # setup a request
print(x.read())                # read page and show the result to the user
```

Python 3 version

10

## We can be slightly fancier in order to handle different formats and the inevitable internet connection errors

```
import urllib.request          # include the urllib.request module

url = "https://www.utexas.edu/"

try:                          # this 'try' statement tells Python that we might expect an error.
    request = urllib.request.urlopen(url) # setup a request
    page = request.read().decode('utf-8') # save the response
    print(page)                # show the result to the user

except urllib.error.URLError:  # handle a page not found error
    print("Could not find page.")
```

→ Run this...

Python 3 version

11

(Heres' the Python 2 version in case you need it)

```
import urllib2                # include the urllib2 module

url = "https://www.utexas.edu/"

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.URLError:       # handle a page not found error
    print("Could not find page.")
```

→ Run this...

Python 2 version

12

→ We just captured the UT web page and printed it out...

```
>>>
<!doctype html>
<html lang="en" dir="ltr">

<head>
....
<meta name="apple-mobile-web-app-title" content="UT Austin" />
<meta name="description" content="The University of Texas at Austin is a bold, ambitious
leader, providing a first-class education and the tools of discovery to more than 51,000
students." />....
```

...and so on, and on, and on...

13

**That was (more or less) 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...**

14

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

```
import urllib.request
pmid = 11237011

# Insert the pmid where the {} are in the following URL:
url = "https://pubmed.ncbi.nlm.nih.gov/?term={0}[uid]&format=pubmed".format(pmid)

try:
    # there might be an error!
    request = urllib.request.urlopen(url)
    page = request.read().decode('utf-8')
    print(page)

except urllib.error.URLError:
    # handle page not found error
    print("Could not connect to Medline!")
```

Python 3 version

15

## 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://pubmed.ncbi.nlm.nih.gov/?term={0}[uid]&format=pubmed".format(pmid)

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

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

Python 2 version

16



**If you run that program, you should get back...**

```
>>>
<!DOCTYPE html>

.....lots of metadata.....
```

```
OWN - NLM
STAT- MEDLINE
DCOM- 20010322
LR - 20210108
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,
MA 02142, USA. lander@genome.wi.mit.edu
```

**the Medline entry for the human  
genome sequence paper**

[and so on]

17

**If you run that program, you should get back...**

```
>>>
<!DOCTYPE html>

.....lots of metadata.....
```

```
OWN - NLM
STAT- MEDLINE
DCOM- 20010322
LR - 20210108
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,
MA 02142, USA. lander@genome.wi.mit.edu
```

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

[and so on]

18

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

```
import urllib.request
pmid = 11237011

# Insert the pmid where the {} are in the following URL:
url = "https://pubmed.ncbi.nlm.nih.gov/?term={0}[uid]&format=pubmed".format(pmid)

try:
    # there might be an error!
    request = urllib.request.urlopen(url)
    page = request.read().decode('utf-8')
    print(page)

except urllib.error.URLError:
    # handle page not found error
    print("Could not connect to Medline!")
```

Python 3 version

19

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

```
import urllib.request
pmid = 11237011

# Insert the pmid where the {} are in the following URL:
url = "https://pubmed.ncbi.nlm.nih.gov/?term={0}[uid]&format=pubmed".format(pmid)

try:
    # there might be an error!
    request = urllib.request.urlopen(url)
    page = request.read().decode('utf-8')
    print(page.count("AU - "))

except urllib.error.URLError:
    # 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**

Python 3 version

20

### (& the Python 2 version, just for the sake of completeness)

```
import urllib2
pmid = 11237011

# Insert the pmid where the {} are in the following URL:
url = "https://pubmed.ncbi.nlm.nih.gov/?term={0}[uid]&format=pubmed".format(pmid)

try:
    request = urllib2.urlopen(url)
    page = request.read()
    print(page.count("AU - "))
except urllib2.URLError:
    print("Could not connect to Medline!")
```

Python 2 version

21

- 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...

22