You and your (DNA) parasites

A retrotransposon someplace in your DNA:

- Assorted genes
- Inverted repeating sequences
- Makes an RNA copy of itself
- Creates a DNA copy of the RNA copy at a new location in your genome (& now you have 2 copies...)

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

---

**Major types of repeats in the human genome**

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

**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
→ 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.

<table>
<thead>
<tr>
<th>Database</th>
<th>Records</th>
<th>Address</th>
</tr>
</thead>
<tbody>
<tr>
<td>BioGRID</td>
<td>&gt;2 M protein interactions</td>
<td><a href="https://thebiogrid.org">https://thebiogrid.org</a></td>
</tr>
<tr>
<td>Ensembl (+ BioMurt for easy sequence queries)</td>
<td>Major repository of DNA sequences, genomes, genes, proteins, and transcripts</td>
<td><a href="http://useast.ensembl.org/index.html">http://useast.ensembl.org/index.html</a></td>
</tr>
<tr>
<td>Expression Atlas</td>
<td>130K mRNA expression expts in 65 species</td>
<td><a href="https://www.ebi.ac.uk/gxa/home">https://www.ebi.ac.uk/gxa/home</a></td>
</tr>
<tr>
<td>Genbank</td>
<td>&gt;1 trillion bases sequenced; &gt;14 trillion bases as whole genome shotgun data</td>
<td><a href="https://www.ncbi.nlm.nih.gov/genbank/">https://www.ncbi.nlm.nih.gov/genbank/</a></td>
</tr>
<tr>
<td>Genomes Online Database (GOLD)</td>
<td>&gt;10K genome sequences, many in progress</td>
<td><a href="https://gold.jgi.doe.gov/index">https://gold.jgi.doe.gov/index</a></td>
</tr>
<tr>
<td>Human Protein Atlas</td>
<td>Millions of high-res images of &gt;17K human proteins across tissues, cancers, &amp; cell lines</td>
<td><a href="http://www.proteinatlas.org/">http://www.proteinatlas.org/</a></td>
</tr>
<tr>
<td>KEGG</td>
<td>Most known pathways, in &gt;5K graphical diagrams and &gt;1K organisms (via homology)</td>
<td><a href="http://www.genome.ad.jp/kegg/">http://www.genome.ad.jp/kegg/</a></td>
</tr>
<tr>
<td>Mouse Genome Informatics</td>
<td>~20,000 mouse genes, diverse associated data &amp; annotations</td>
<td><a href="http://www.informatics.jax.org/">http://www.informatics.jax.org/</a></td>
</tr>
<tr>
<td>Pride</td>
<td>Hundreds of millions of peptide mass spectra from &gt;200 of thousands of experiments</td>
<td><a href="https://www.ebi.ac.uk/pride/archive/">https://www.ebi.ac.uk/pride/archive/</a></td>
</tr>
<tr>
<td>Reactome</td>
<td>&gt;2K pathways involving &gt;10K human proteins, also other organisms</td>
<td><a href="https://www.reactome.org/">https://www.reactome.org/</a></td>
</tr>
<tr>
<td>SGD</td>
<td>~6,000 yeast genes, diverse associated data &amp; annotations</td>
<td><a href="https://www.yeastgenome.org/">https://www.yeastgenome.org/</a></td>
</tr>
<tr>
<td>UniProtKB/SWISS-PROT</td>
<td>&gt;550K hand-curated sequence entries from &gt;14K organisms</td>
<td><a href="https://www.uniprot.org/">https://www.uniprot.org/</a></td>
</tr>
<tr>
<td>Database</td>
<td>Record</td>
<td></td>
</tr>
<tr>
<td>-------------------------------</td>
<td>---------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>BioGRID</td>
<td>&gt;2 M protein-protein interactions</td>
<td></td>
</tr>
<tr>
<td>EcoCyc/MetaCyc</td>
<td>&gt;2,700 pathways from &gt;130 organisms</td>
<td></td>
</tr>
<tr>
<td>Ensembl (+ BioMart for easy sequence queries)</td>
<td>Major repository of DNA sequences, genes, proteins, and transcripts</td>
<td></td>
</tr>
<tr>
<td>Entrez Genome</td>
<td>Thousands of genome sequences</td>
<td></td>
</tr>
<tr>
<td>Expression Atlas</td>
<td>123K mRNA expression certs in 62 species</td>
<td></td>
</tr>
<tr>
<td>Genbank</td>
<td>&gt;300 M bases sequenced; &gt;5 trillion bases as whole genome shotgun data</td>
<td></td>
</tr>
<tr>
<td>Gene Expression Omnibus (GEO)</td>
<td>&gt;4 M mRNA or protein expression results</td>
<td></td>
</tr>
<tr>
<td>Genomes Online Database (GOLD)</td>
<td>&gt;150K genome sequences, many in progress</td>
<td></td>
</tr>
<tr>
<td>Human Protein Atlas</td>
<td></td>
<td></td>
</tr>
<tr>
<td>KEGG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medline / PubMed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mouse Genome Informatics</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Online Mendelian Inheritance in Man (OMIM)</td>
<td>Compendium of human genes and genetic phenotypes, data for &gt;16,000 human genes</td>
<td></td>
</tr>
<tr>
<td>Pride</td>
<td>Hundreds from 19P</td>
<td></td>
</tr>
<tr>
<td>Reactome</td>
<td>&gt;2K pathways also other</td>
<td></td>
</tr>
<tr>
<td>SGD</td>
<td>&gt;3000 genes</td>
<td></td>
</tr>
<tr>
<td>UniProtKB/SWISS-PROT</td>
<td>&gt;550K hand-curated sequence entries from &gt;9K organisms</td>
<td></td>
</tr>
</tbody>
</table>

Just some of the resources available for bioinformatics

Think of these as the raw data for new discoveries

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

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

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](https://docs.python.org/3.8/library/urllib.request.html) or [http://docs.python.org/2/library/urllib2.html](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:

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

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

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

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

→ Run this...
```

Python 3 version

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

```python
import urllib2  # include the urllib2 module

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

try:
    request = urllib2.urlopen(url)  # setup a request
    page = request.read()  # save the response
    print(page)  # show the result to the user
except urllib2.URLError:
    print("Could not find page.")  # handle a page not found error

→ Run this...
```

Python 2 version
We just captured the UT web page and printed it out...

```html
<!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...
</head>
```

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

```python
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:
    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

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

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

Python 2 version
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
[and so on]
```

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
[and so on]
```

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

```python
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

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

Run this, & get ...

```bash
>>> 256
```

So, there were 256 authors on one (of the two) human genome papers
(& the Python 2 version, just for the sake of completeness)

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

Python 2 version

• Queries to Medline or any other NCBI database, including GenBank, are described at: [http://www.ncbi.nlm.nih.gov/books/NBK3862/](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...