## Science news of the day:



**CROWDSOURCING**

**Online Video Game Plugs Players Into Remote-Controlled Biochemistry Lab**

*Science* 343 (2014)

"The game, called EteRNA, allows players to remotely carry out real experiments to verify their predictions of how RNA molecules fold."

"hooking players up directly to a real-world, robot-controlled biochemistry lab."

In tests, … "human players still came out on top, solving structures more accurately than the standard software 99% of the time."

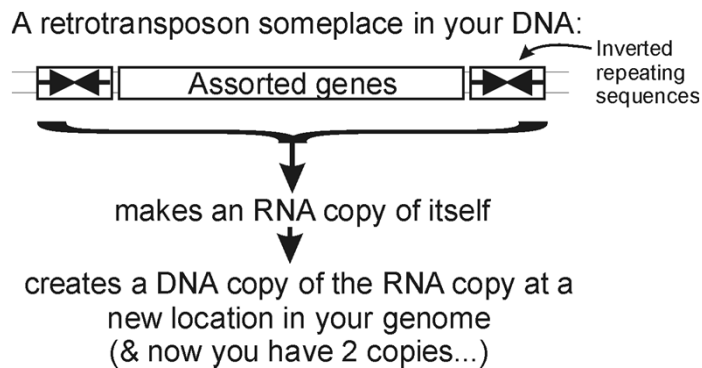**The paper reporting EteRNA (Lee et al., PNAS, 2014) has >37,000 authors.**

Edward Marcotte/Univ. of Texas/BIO337/Spring 2014

---

## Wrapping up our last topic:
## 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…**

Edward Marcotte/Univ. of Texas/BIO337/Spring 2014

**Wrapping up our last topic:**
**You and your (DNA) parasites**

Classes of interspersed repeat in the human genome

|  |  |  | Length | Copy number | 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.**

Edward Marcotte/Univ. of Texas/BIO337/Spring 2014 (apologies—missing the citation, now lost)

---

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

Edward Marcotte/Univ. of Texas/BIO337/Spring 2014

## Slide 1

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

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

| Database | Records | Address |
|---|---|---|
| dbEST | 74,186,692 public entries | http://www.ncbi.nlm.nih.gov/dbEST/ |
| DIP | 75,019 protein interactions | http://dip.doe-mbi.ucla.edu/ |
| EcoCyc/MetaCyc | >1,900 pathways | http://www.ecocyc.org, http://www.metacyc.org |
| Entrez Genome | 1000's of genomes (including ~4,500 viruses) | http://www.ncbi.nlm.nih.gov/genome?db=genome |
| Genbank | 135,440,924 sequence records spanning 126 billion bases in traditional Genbank (as of 2013); 191 billion bases in WGS division | http://www.ncbi.nlm.nih.gov/Genbank/GenbankOverview.html |
| Gene Expression Omnibus (GEO) | 877,498 mRNA or protein expression data sets | http://www.ncbi.nlm.nih.gov/geo/ |
| Genomes Online Database (GOLD) | 20,581 genome sequences (many in progress) | http://www.genomesonline.org/cgi-bin/GOLD/index.cgi |
| Human Protein Atlas | millions of images of ~14K human proteins' expression in 46 tissues, 20 cancers, 47 cell lines | http://www.proteinatlas.org/ |
| KEGG | Most known pathways, in 435 graphical diagrams and 2,455 organisms (via homology) | http://www.genome.ad.jp/kegg/ |
| Medline | >22 million references | http://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 >12,000 genes | http://www.ncbi.nlm.nih.gov/omim/ |
| Pride | > 342 million peptide mass spectra from 27K experiments | http://www.ebi.ac.uk/pride/ |
| Reactome | 1,371 pathways involving 6,571 proteins, for human, similar for extra organisms | http://www.reactome.org/ |
| SGD | ~6,000 yeast genes, diverse associated data & annotations | http://www.yeastgenome.org/ |
| Yeast GFP database | protein subcellular localization for ~4,500 yeast proteins | http://yeastgfp.yeastgenome.org/ |
| Yeast regulatory network | ~11,000 transcription factor/downstream gene pairs | http://web.wi.mit.edu/young/regulatory_code/ |

Edward Marcotte/Univ. of Texas/BIO337/Spring 2014

## Slide 2

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

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

| Database | Records | |
|---|---|---|
| dbEST | 74,186,692 public entries | dbEST/ |
| DIP | 75,019 protein interactions | |
| EcoCyc/MetaCyc | >1,900 pathways | http://www.ecocyc.org, http://www.metacyc.org |
| Entrez Genome | 1000's of genomes (including ~4,500 viruses) | |
| Genbank | 135,440,924 sequence records spanning 126 billion bases in traditional Genbank (as of 2013); 191 billion bases in WGS division | |
| Gene Expression Omnibus (GEO) | 877,498 mRNA or protein expression data sets | |
| Genomes Online Database (GOLD) | 20,581 genome sequences (many in progress) | |
| Human Protein Atlas | milli... expr... | http://www.proteinatlas.org/ |
| KEGG | Most... and 2... | |
| Medline | >22... | |
| Mouse Genome Informatics | ~20,000 mouse genes, diverse associated data & annotations | |
| Online Mendelian Inheritance in Man (OMIM) | Compendium of human genes and genetic phenotypes, data for >12,000 genes | |
| Pride | > 342 million peptide mass spectra from 27K experiments | |
| Reactome | 1,371 pathways involving 6,571 proteins, for human, similar for extra organisms | |
| SGD | ~6,000 yeast genes, diverse associated data & annotations | |
| Yeast GFP database | protein subcellular localization for ~4,500 yeast proteins | http://yeastgfp.yeastgenome.org/ |
| Yeast regulatory network | ~11,000 transcription factor/downstream gene pairs | http://web.wi.mit.edu/young/regulatory_code/ |

Annotations:
- >75K protein-protein interactions
- GEO has ~900K experiments, each measuring 1000's of mRNA or protein abundances
- Medline has >22 million research articles, many with complete text online
- OMIM = the most important resource for human genetic disease
- >1,300 biochemical processes and reactions, described in detail

Edward Marcotte/Univ. of Texas/BIO337/Spring 2014

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….**

We saw one way to do this in AppSoma.  Here's another.

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

→ **Run this…**

---

→ **We just captured the UT web page and printed it out (minus the images)…**

```
>>>
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN"
  "http://www.w3.org/TR/xhtml1/DTD/xhtml1-strict.dtd">
<html xmlns="http://www.w3.org/1999/xhtml" xml:lang="en" lang="en" dir="ltr">

<head>
<meta http-equiv="Content-Type" content="text/html; charset=utf-8" />
  <meta http-equiv="Content-Type" content="text/html; charset=utf-8" />
<link rel="shortcut icon" href="/sites/default/files/webcentral_favicon_0.ico"
type="image/x-icon" />
  <title>Home | The University of Texas at Austin</title>
  <link type="text/css" rel="stylesheet" media="all"
href="/sites/default/files/css/css_39da96d3505397e78e1b5ff4743df898.css" />
  <script type="text/javascript"
src="/sites/default/files/js/js_bb53cbf52ebfa96587a1c184036c356e.jsmin.js"></script>
<script type="text/javascript">
<!--//--><![CDATA[//><!—
                                    …and so 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:**

```python
import urllib2
pmid = 11237011

# Insert the pmid where the {} are in the following URL:
url = "http://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**

Edward Marcotte/Univ. of Texas/BIO337/Spring 2014

---

**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…**

Edward Marcotte/Univ. of Texas/BIO337/Spring 2014

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

```python
import urllib2
pmid = 11237011

# Insert the pmid where the {} are in the following URL:
url = "http://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?**

```python
import urllib2
pmid = 11237011

# Insert the pmid where the {} are in the following URL:
url = "http://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 ...**   >>>
                              255

> **So, there were 255 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/

- 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...