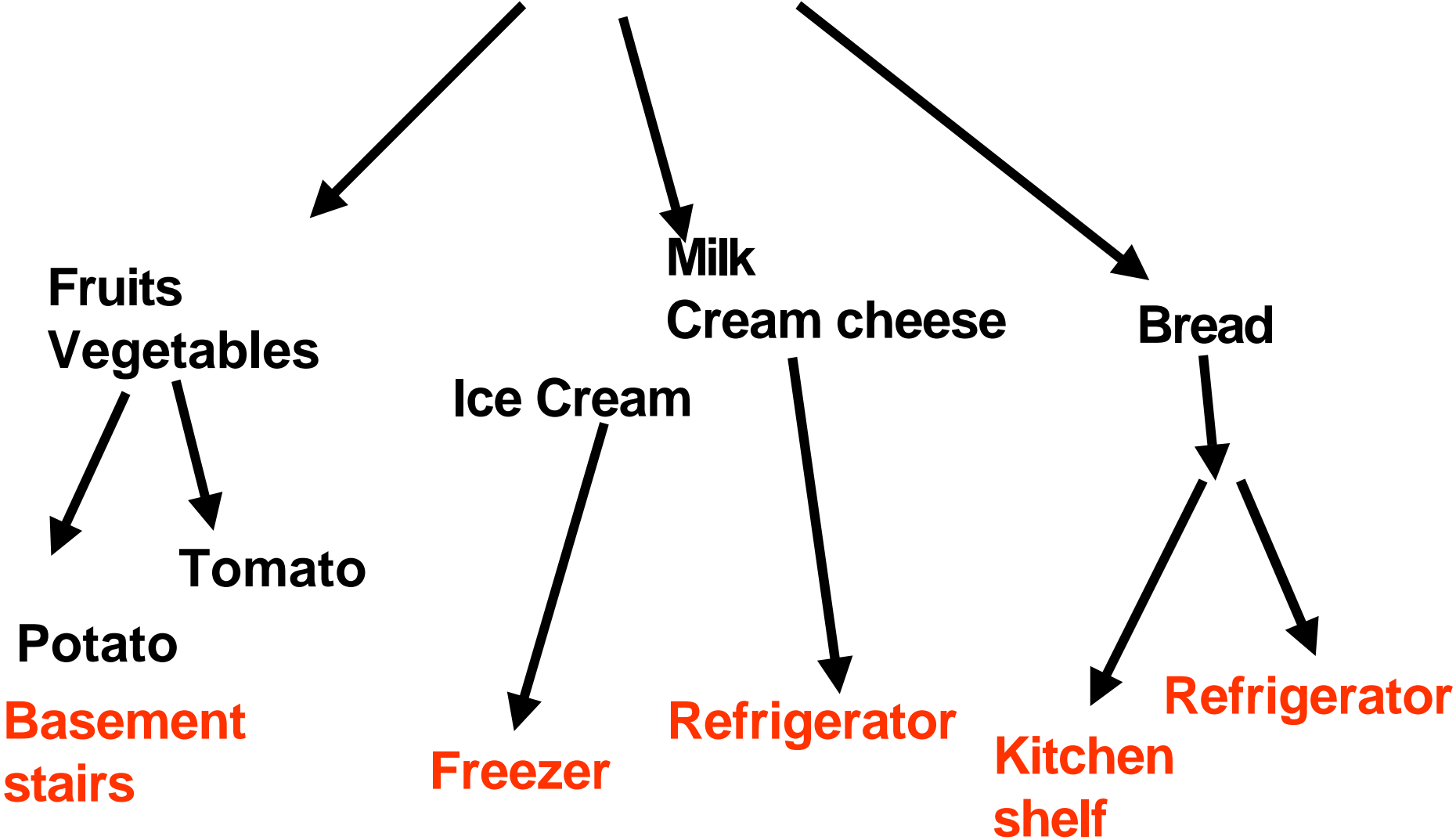


# More on Parsing Data Files

## Bag of Grocery Stuff



## Let's have a blast

```
from Bio import Fasta
```

```
file_for_blast = open('m_cold.fasta', 'r')
```

```
f_iterator = Fasta.Iterator(file_for_blast)
```

```
f_record = f_iterator.next()
```

```
from Bio.Blast import NCBIWWW
```

```
b_results = NCBIWWW.blast('blastn', 'nr', f_record)
```

```
save_file = open('my_blast.out', 'w')
```

```
blast_results = b_results.read()
```

```
save_file.write(blast_results)
```

```
save_file.close()
```

## Let's go to the Pub Med

```
from Bio.Medline import PubMed  
from Bio import Medline
```

```
search_term = 'Sxl'  
search_ids = PubMed.search_for(search_term)
```

```
print 'search_ids:\n', search_ids
```

**search\_ids:**

**['11910129', '11565743', '11309853', '11238934', '11238895', '11222155', '11195332', '11156988', '11156616', '11118879', '10924474', '10921779', '10882142', '10879541', '10864047', '10835396', '10791828', '10790389', '10780450', '10761105', '10734193', '10631327', '10617208', '10588726', '10572053', '10556072', '10555143', '10545124', '10521666', '10393111', '10357929', '10220389', '10217141', '10102992', '10101174', '10082569', '10071220', '9883585', '9671597', '9642170', '9502731', '9502730', '9502729', '9592147', '9570314', '9398148', '9363684', '9362474', '9299339', '9268369', '9207093', '9144292', '9096138', '9092663', '9032294', '8978052', '9003283', '8982872', '8913748', '8787749', '8756662', '8817458', '8631274', '8846292', '8601486', '8575302', '7563134', '7556096', '7567454', '7623836', '7651341', '7796814', '7761834', '7823955', '7768187', '7760212', '7713421', '7867511', '7524663', '7524034', '7985243', '8069866', '7958879', '8062831', '7811638', '7516476', '7926760', '8162863', '8013251', '8005414', '8288132', '8062458', '8062457', '7835150', '8246990', '8370520', '8275850', '8076519', '8076518', '7680770', '8330539', '8330537', '8422681', '1454517', '1525829', '1511868', '1592233', '1547493', '1551580', '1748277', '1782877', '1743482', '1710769', '2015624', '1900493', '1904047', '2054887', '2124516', '1695150', '2193725', '1690860', '2612376', '2513254', '2583094', '2473007', '2702687', '2924997', '3144435', '3145197', '3137120', '2840023', '3129196', '2822534', '3802198', '2903043', '3030733', '3089868', '3453784', '3000609', '3931920', '3920120', '6402396', '6785042', '6790338', '7357715', '105964']**

PubMed continued...

```
import string
```

```
rec_parser = Medline.RecordParser( )
```

```
medline_dict = PubMed.Dictionary(parser = rec_parser)
```

```
for id in search_ids[0:5]:
```

```
    cur_record = medline_dict[id]
```

```
    print 'title:', string.rstrip(cur_record.title)
```

```
    print 'authors:', cur_record.authors
```

```
    print 'source:', string.strip(cur_record.source)
```

```
    print
```

**title: The Drosophila fl(2)d gene, required for female-specific splicing of Sxl and tra pre-mRNAs, encodes a novel nuclear protein with a HQ-rich domain.**

**authors: ['Penalva LO', 'Ruiz MF', 'Ortega A', 'Granadino B', 'Vicente L', 'Segarra C', 'Valcarcel J', 'Sanchez L']**

**source: Genetics 2000 May;155(1):129-39.**

**Are**  
**we**  
**all**  
**aligned ?**

↓ **Clustalw**

	<b>Y</b>		<b>es</b>
<b>Now</b>			
<b>w</b>			<b>e</b>
<b>a</b>	<b>ll</b>		
<b>a</b>		<b>r</b>	<b>e</b>
<b>a</b>	<b>l</b>	<b>igned</b>	

## Alignment

```
import os  
import sys
```

```
from Bio.Clustalw import MultipleAlignCL  
from Bio import Clustalw  
from Bio.Align import AlignInfo
```

```
cline = MultipleAlignCL(os.path.join(os.curdir,  
                                     'opuntia.fasta'))
```

```
cline.set_output('opuntia.aln')
```

Alignment continued...

```
alignment = Clustalw.do_alignment(cline)
all_records = alignment.get_all_seqs()

print 'description:', all_records[0].description
print 'sequence:', all_records[0].seq
print "alignment:", alignment
length = alignment.get_alignment_length()
print "length: ", length

summary_align = AlignInfo.SummaryInfo(alignment)

consensus = summary_align.dumb_consensus()
print "consensus:", consensus
```

...

gi		6273285		gb		AF191659.1		AF191	TATATA-----ATATATTTCAAATTTCCTTATATACCCA
gi		6273284		gb		AF191658.1		AF191	TATATATA-----ATATATTTCAAATTTCCTTATATACCCA
gi		6273287		gb		AF191661.1		AF191	TATATA-----ATATATTTCAAATTTCCTTATATATCCA
gi		6273286		gb		AF191660.1		AF191	TATATA-----ATATATTTATAATTTCCTTATATATCCA
gi		6273290		gb		AF191664.1		AF191	TATATATATA-----ATATATTTCAAATTTCCCTTATATATCCA
gi		6273289		gb		AF191663.1		AF191	TATATATATA-----ATATATTTCAAATTTCCCTTATATATCCA
gi		6273291		gb		AF191665.1		AF191	TATATATATATATATAATATATTTCAAATTTCCCTTATATATCCA

\*\*\*\*\*                   \*\*\*\*\*       \*\*\*\* \*\*\*\*\*       \*\*\*

...

length:  
906

consensus:

Seq ( 'TATACATTAAAGNAGGGGGATGCGGATAAATGGAA ...',  
IUPACAmbiguousDNA ( ) )

**Bye for now!!!**  
**See you next week at the**

**GenBank**

**party**