MCB 372

Student Projects

Databanks, Blast

possibly unix Perl

J. Peter Gogarten
Office: *BPB 404*
phone: *860 486-4061,*
Email: *gogarten@uconn.edu*

Student Projects

- Should be related to your interests !!!
- Examples for possible projects:

Example 1: Evolution of a gene family

- When in the evolution of the interferon (or what ever you are interested in) gene family did gene duplications occur?
- Which of the resulting subfamilies have acquired a new function?
- What is the phylogenetic distribution of this subfamily? (Would you expect members of this subfamily to be present in insects, fish, chicken, fungi, archaea?)
- Can you detect episodes of positive selection?
- Is there anything that would suggest gene conversion events?

The “ to-do-list” would include:

- gather data (note for some of the questions mentioned above you'll need aa **and** nucleotide sequences),
- align sequences
- build phylogenies
- analyze sequences
- assess reliability of branches
- INTERPRET WHAT YOU GOT!

Example 2: Can one detect a distinct second peak in the divergence of putatively chimeric genomes?

Genome fusions are the latest rage in evolutionary biology:
For example:

- Koonin EV, Mushegian AR, Galperin MY, Walker DR. *Comparison of archaeal and bacterial genomes: computer analysis of protein sequences predicts novel functions and suggests a chimeric origin for the archaea.* Mol Microbiol. 1997 Aug;25(4):619-37.
- The Eukaryotes are a chimera of at least an archaeal like host cell and a bacterium that evolved into a mitochondrium (+ in some cases a cyanobacterium that evolved into a plastid)
- The Haloarchaea contain many bacterial genes
- The Thermotogales contain many archaeal genes
- Most plants and many fungi (likely including bakers yeast) are aneupolyploids

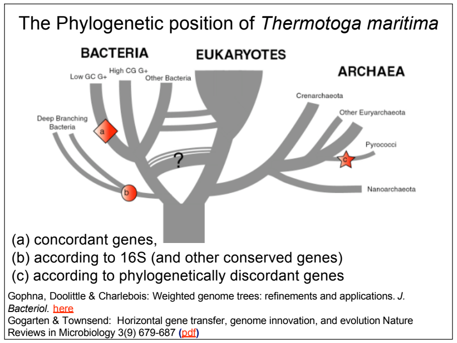
In most of these instances it is not clear that the transfer (duplication) really occurred in a single massive event, or if the transfers (duplications) occurred on a gene by gene basis.
(in yeast the type of genes that were duplicated suggest distinct selection pressures, see Benner et al here)

Example 2: Chimera? continued

In case of a chimera formed in a single historic event one would expect

A) Two distinct types of phylogenetic affinity.

E.g.: Genes in Thermotoga maritima should either group with the sistergroup of the bacterial partner, or with the sistergroup of the archaeal donor



The Phylogenetic position of *Thermotoga maritima*

(a) concordant genes,
(b) according to 16S (and other conserved genes)
(c) according to phylogenetically discordant genes

Gophna, Doolittle & Charlebois: Weighted genome trees: refinements and applications. *J. Bacteriol.* here
Gogarten & Townsend: Horizontal gene transfer, genome innovation, and evolution Nature Reviews in Microbiology 3(9) 679-687 (pdf)

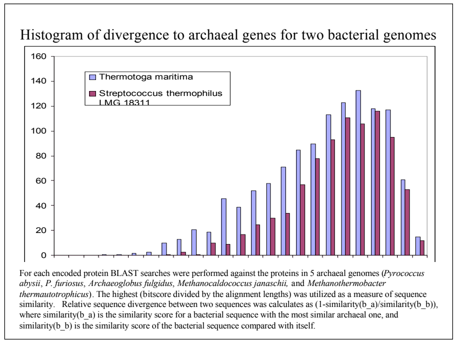
Example 2: Chimera? continued

In case of a chimera formed in a single historic event one would expect

A) Two distinct types of phylogenetic affinity.

E.g.: Genes in Thermotoga maritima should either group with the sistergroup of the bacterial partner, or with the sistergroup of the archaeal donor

B) Two distinct peaks in a divergence histogram.
E.g.: If one measures the divergence Thermotoga – Archaea for all the individual genes, under the assumption of a chimera formation one should obtain a bimodal distribution in a histogram of the different genes.



Histogram of divergence to archaeal genes for two bacterial genomes

For each encoded protein BLAST searches were performed against the proteins in 5 archaeal genomes (*Pyrococcus abysii*, *P. furiosus*, *Archaeoglobus fulgidus*, *Methanocaldococcus janaschii,* and *Methanothermobacter thermautotrophicus*). The highest (bitscore divided by the alignment lengths) was utilized as a measure of sequence similarity. Relative sequence divergence between two sequences was calculates as (1-similarity(b_a)/similarity(b_b)), where similarity(b_a) is the similarity score for a bacterial sequence with the most similar archaeal one, and similarity(b_b) is the similarity score of the bacterial sequence compared with itself.

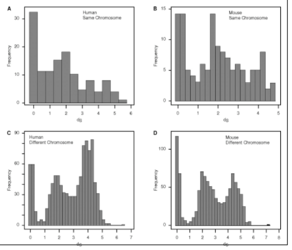
Example 2, continued

The “ to-do-list” would include:

- Formulate the question you want to address
- Find a computer where you can run blastall (this might take a couple of hours)
- Download and analyze the required genomes
- Analyze the results in an Excel spreadsheet
- Selected some genes (e.g., the ones that are most archaeal), assemble gene families and reconstruct their phylogenies.
- INTERPRET YOUR RESULTS! What does it all mean?

Example 3: Gene versus Genome Duplications

The same approach as suggested for the chimera formation can be applied to the question was the whole genome or a large segment of an organism’s genome duplicated, or did the duplications occur in a piecemeal fashion?



Frequency distributions of d_S in human and mouse between the members of two-member gene families located on the same and different chromosomes

From: Robert Friedman and Austin L. Hughes: *Two Patterns of Genome Organization in Mammals: the Chromosomal Distribution of Duplicate Genes in Human and Mouse.* Mol. Biol. Evol. 21(6):1008–1013. 2004

Background for Example 3:
Selection acts on

- **genes** (as in the selfish gene theory, the genes are the replicators that build the body of the organism). According to this all genes are selfish, most are cooperating with one another, a few are not. To distinguish the latter from the former, I call them parasitic genes (or molecular parasites).
- **individuals** in a population (the survival of the fittest).
- **groups** of organisms (group selection). The group that has properties that allows it to adapt better, or to evolve faster, or to make better use of resources will be selected. In this case the group (community, *not necessarily all belonging to the same species*) is the unit of selection. (see group selection entry at wikipedia)

Note: in general this is controversial. To what extent is group selection reflecting kin-selection: the organism acting to guarantee survival of genes that are related to its own genes (bees in a beehive are all closely related).

Examples for “group selection” in microbes: (a) *Agrobacteria*

Agrobacteria that carry a Ti plasmid can transform plant cells with a T DNA. As result of a successful transformation the plant cell has integrated the T DNA into its genome and expresses the encoded genes. This results in the transformed cells forming a tumor, and, in addition, the transformed plant cells also produce a strange amino acid that cannot be utilized by the plant cells, but that serves as a carbon and nitrogen source for the *Agrobacteria*. The genes responsible for transferring the Ti plasmid between different *Agrobacteria* (*tra* genes) are under the control of quorum sensing. The effect is that if one *Agrobacterium* strain has successfully transformed a plant, and now lives from the plant produced strange amino acid, other *Agrobacteria* can receive the Ti plasmid, which contains the T DNA transferred into the plant and in addition encodes enzymes that allow the metabolism of the strange amino acids. The *Agrobacteria*, which receive the Ti-plasmid thus participate in the utilization of the plant produced carbon and nitrogen source. This observation **might be described as group selection**: the population of *Agrobacteria* avoids a selective sweep and carries larger genetic diversity into the population living on the transformed plant. The increased diversity will facilitate future adaptations to a changing environment, and will avoid the fixation of slightly deleterious mutations that might have been carried by the *Agrobacterium* that transformed the plant cell. On the other hand, **one can consider this process the outcome of the "selfishness" of the *tra*-genes and of the Ti plasmid**. These genes manage to move themselves into the growing part of the population, and they will benefit form a more diverse group of host organisms.

Examples for “group selection” in microbes (b):
Metal resistance genes in microbial communities inside rocks in the dry valleys of Antarctica

These rocks have high concentrations of toxic heavy metals. The endolithic microbial community readily shares heavy metal resistant genes with microbes that might be able to become part of the community. At the community level the outcome is a higher diversity, and a richer network of metabolic reactions. Presumably the more diverse communities are more stable towards perturbations, and provided the community can propagate as a whole, this would provide a **selective advantage to the community**. However from the **selfish gene point of view**, the resistance gene increases its chances of long term survival by invading as many additional species as possible.

Examples for “group selection” in microbes (c):
Gene Transfer Agents (GTA) in alpha proteobacteria

GTAs are propahges that do not specifically pack their own DNA, but that unselectively pack host DNA into the phage head (see here).

• Are these just defective prophages that lost their sequence specificity in DNA packaging?
• Is this an illustration that HGT is beneficial and under group selection?

(Aside: In general, HGT might reflect uptake of DNA for food, recombination might be a negligible side effect (Rosi Redfield, e.g. here), or HGT might reflect the selfishness of the transferred DNA.

Testing GTAs as agents selected by group selection.

Possible hypotheses:
- GTAs are defective prophages that lost their sequence specificity in DNA packaging?
- GTAs evolved from phages but now benefit the group and are under group selection?

Under #2:
- The GTA should be more related to one another than to functioning phage
- There molecular phylogeny should reflect reflect the phylogeny of the organism (as measured by rRNA and ribosomal proteins
- The genes encoding the GTA should be under strong purifying selection (under #1 they should be psudogenes).

GTAs: to do list

❖ Identify GTAs in genomes of closely related organisms.
❖ Align the major conserved genes from these GTAs.
❖ Include an appropriate outgroup
From the same genome select genes from the translation machinery, whose phylogeny likely reflects the main current of the organismal history.
❖ Calculate and compare the phylogenies.
❖ Test the GTA genes for positive/purifying selection

other ideas:

➢ Write a script that uses the 100+ known intein alleles each as a seed in PSI BLAST, and stores the profiles. Write a second script that uses these profiles to detect putative inteins in completely sequenced genomes.
➢ Same as above but use transposases, integrases, homing endonucleases, or a molecular parasite of your choice as a seed.
➢ Determine the impact of HGT on reconstruction of organismal evolution. Use one of the several available programs to simulate sequence evolution for several genes along a tree. Reconstruct the phylogeny using either the concatenated genes, or the individual data sets (in the latter case use a super tree approach to calculate the organismal tree as consensus.
Which approach (supertree versus concatenation) recovers the correct tree?
Use different approaches to identify the transferred genes.
➢ Search the different versions of the Mosquito genome for genes from Aeromonas.
➢ Form families for all genes from Thermotogales, add the fifteen most similar sequences from reference genomes, calculate phylogenies, screen for polyphyly of Thermotogales, screen for conflict with consensus.

Assignments for next week:

Think about a topic for your student project!
Please, don’t hesitate to send me an email in case you have a question.

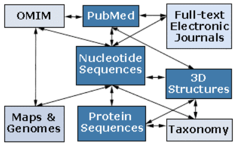
Let me know what you are interested in (email).
What we will do in this course will in part depend on your interests.

Reading for Monday:
Read through the NCBI's BLAST tutorial

Databanks (A)

NCBI (National Center for Biotechnology Information) is a home for many public biological databases (see an older diagram below). All of the databases are interlinked, and they all have common search and retrieval system - **Entrez**.

Another more complete representation with an interactive display of the number of the connections between the different databases in ENTRZ is here.



Entrez / Pubmed, continued

- An interactive Pubmed tutorial click here.
- An Entrez tutorial (non interactive) is here
- Use Boolean operators (**AND**, **OR**, **NOT**) to perform advanced searches. Here is an explanation of the Boolean operators from the Library of Congress Help Page.

- Explore features of Entrez interface: **Limits**, **Index**, **History**, and **Clipboard**.
- Search Field Tags- Listed here.

Other Literature databanks and Services

While Pubmed is incorporating more and more non-medical literature, there might still be gaps in the coverage. Alternatives are local services offered at the UConn libraries. Especially Current Contents and Agricola nicely complement PubMed. The best way to access them is the use of "SilverPlatter" database.

Also, the "Web of Science" database gives access to the Science Citation Index: a database that tracks cited references in journals. Note that these resources are restricted to UConn domain, so you either need to access it from a campus computer or through the proxy account.

Search Robots

PubCrawler allows to run predefined literature searches. Results are written into a database and you are send an email, if there were new results. NCBI now offers a similar service (see My NCBI (Chubby), check the tutorial).

Swiss-Shop is offering the same service for proteins

Sequence and structure databanks

can be divided into many different categories.
One of the most important is

Supervised databanks with gatekeeper. Examples:	Repositories without gatekeeper. Examples:
Swissprot	GenBank
Refseq (at NCBI)	EMBL
	TrEMBL
Entries are checked for accuracy.	Everything is accepted
+ more reliable annotations	+ everything is availabel
-- frequently out of date	-- many duplicates
	-- poor reliability of annotations

Other web pages besides the NCBI

•*Nucleic Acid Research Database Issue* Every year, the first issue of *Nucleic Acid Research* is devoted to updates on biological databases.

•http://www.ebi.ac.uk/ The European homolog/analog to NCBI.

•http://rdp.cme.msu.edu/ The US ribosomal databank project

•http://www.jgi.doe.gov/ The Joint Genome Institute
A recent addition is the integrated microbial genomes site at http://img.jgi.doe.gov/, the coolest feature is the selected gene neighborhoods.

•http://www.genomesonline.org/ Most up to date information on ongoing and completed genome projects – free for academic users.

Several more organism specific resources:
•*http://genome-www.stanford.edu/ Yeast and Arabidopsis genome projects*
•*http://www.flybase.org/ Database of Drosophila Genome*
•*http://www.arabidopsis.org/ TAIR - The Arabidopsis Information Resource*
• *http://www.ensembl.org/ Ensembl Genome Browser (Eukaryotic genomes, including Human and Mouse genomes)*

UNIX

Basic UNIX commands
ls, cd, chmod, cp, rm, mkdir, more (or) less, vi, ps, kill −9, man
A brief listing is here

chmod is a particular pain in the ... .
Under unix every file has an owner and the owner, his group and everyone else have permissions to read, write and/or execute the file (or they don't). If you want to see which permissions are currently assigned to your files, type ls -l at the command prompt.
chmod a+x *.pl gives everyone execute permission for all files that end with .pl the * is a wildcard. (warning don't ever use rm in conjunction with *)
For more on chmod type "man chmod" or see here.
(In the OSX GUI you can control click at a file, and change permissions in the info box). Most ssh clients (FUGU and SSH) allow you to use a GUI to change file permissions (in FUGU ctrl click).

Unix - command line interface

If you tried to execute a command, and you made a mistake, for example, you mistyped a file name, you can recall the last command using the up arrow (down arrow for more recent).

If you are tired typing long filenames, you can use the tab key to complete the line, provided there is only one way to complete the line. E.g: cd /Desktop could be replaced by cd /D<tab>
If there are two or more choices you hear a boing, if you hit <tab> again, you get a list of choices.

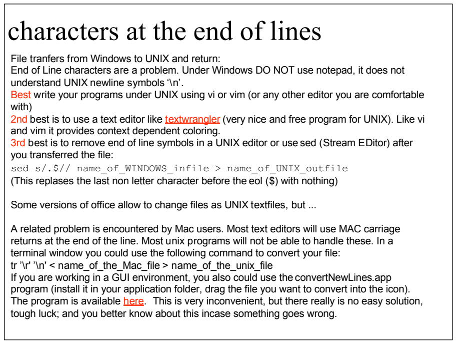
writing Perl scripts

Use unix/ linux /OsX if possible (talk with Tim if you want to use windows).
A) open a terminal window ; type "which perl <return>"
B) SSH to a unix machine (cluster OsX), log in, type "which perl <return>"
C) to check the version type perl -v <return>The response of the system should tell you, where Perl is installed on your machine (you need to know this for the first line of your perl program, which tells the operating system how to interpret what follows. On most installations this is #!/usr/bin/perl).

WINDOWS: If you use a windows machine, you can use an ssh program to connect to the biotech cluster. A good ssh client is available at ftp://ftp.ssh.com/pub/ssh/- highly recommended. I am sure that there are editors available that are more useful than notepad, but I don't know of them. :(

MAC OsX: If you use a Mac under OS X, and you do not want to (only) use the PERL locally, you want to install both jellyfish (ssh terminal) and fugu (a secure file transfer program). Both are available at ftp://ftp.uconn.edu/pub/packages/ssh/mac/ or through the people who wrote the software - GOOGLE) Also, the bbcxsrv1 is available as a server using ssh or apl. You can connect to to it from the finder menu (-> GO -> Connect to Server) pasting the following into the menu box afp://bbcxsrv1.biotech.uconn.edu (select your account).

LINUX: Most editors on linux systems recognize Perl programs and provide context dependent coloring. Ssh and Konquerer work well for file transfer.



characters at the end of lines

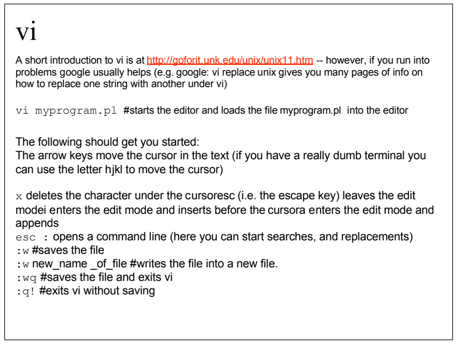
File tranfers from Windows to UNIX and return:
End of Line characters are a problem. Under Windows DO NOT use notepad, it does not understand UNIX newline symbols '\n'.
Best write your programs under UNIX using vi or vim (or any other editor you are comfortable with)
2nd best is to use a text editor like textwrangler (very nice and free program for UNIX). Like vi and vim it provides context dependent coloring.
3rd best is to remove end of line symbols in a UNIX editor or use sed (Stream EDitor) after you transferred the file:

```
sed s/.$// name_of_WINDOWS_infile > name_of_UNIX_outfile
```

(This replases the last non letter character before the eol ($) with nothing)

Some versions of office allow to change files as UNIX textfiles, but ...

A related problem is encountered by Mac users. Most text editors will use MAC carriage returns at the end of the line. Most unix programs will not be able to handle these. In a terminal window you could use the following command to convert your file:
tr '\r' '\n' < name_of_the_Mac_file > name_of_the_unix_file
If you are working in a GUI environment, you also could use the convertNewLines.app program (install it in your application folder, drag the file you want to convert into the icon). The program is available here. This is very inconvenient, but there really is no easy solution, tough luck; and you better know about this incase something goes wrong.



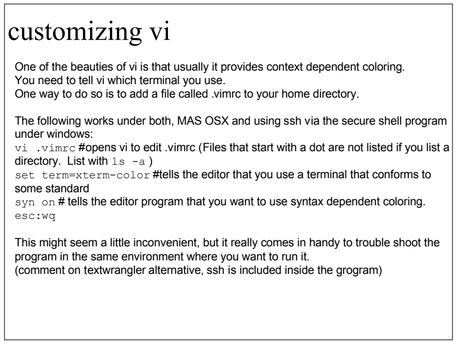
vi

A short introduction to vi is at http://goforit.unk.edu/unix/unix11.htm -- however, if you run into problems google usually helps (e.g. google: vi replace unix gives you many pages of info on how to replace one string with another under vi)

`vi myprogram.pl` #starts the editor and loads the file myprogram.pl into the editor

The following should get you started:
The arrow keys move the cursor in the text (if you have a really dumb terminal you can use the letter hjkl to move the cursor)

`x` deletes the character under the cursoresc (i.e. the escape key) leaves the edit modei enters the edit mode and inserts before the cursora enters the edit mode and appends
`esc :` opens a command line (here you can start searches, and replacements)
`:w` #saves the file
`:w` new_name _of_file #writes the file into a new file.
`:wq` #saves the file and exits vi
`:q!` #exits vi without saving

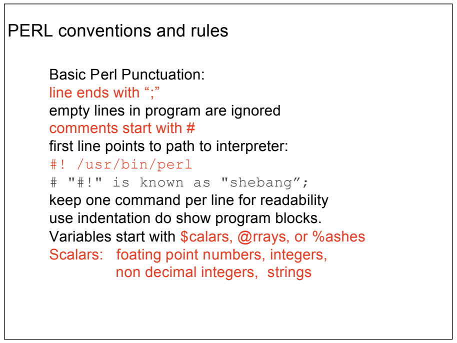


customizing vi

One of the beauties of vi is that usually it provides context dependent coloring.
You need to tell vi which terminal you use.
One way to do so is to add a file called .vimrc to your home directory.

The following works under both, MAS OSX and using ssh via the secure shell program under windows:
`vi .vimrc` #opens vi to edit .vimrc (Files that start with a dot are not listed if you list a directory. List with `ls -a`)
`set term=xterm-color` #tells the editor that you use a terminal that conforms to some standard
`syn on` # tells the editor program that you want to use syntax dependent coloring.
`esc:wq`

This might seem a little inconvenient, but it really comes in handy to trouble shoot the program in the same environment where you want to run it.
(comment on textwrangler alternative, ssh is included inside the grogram)

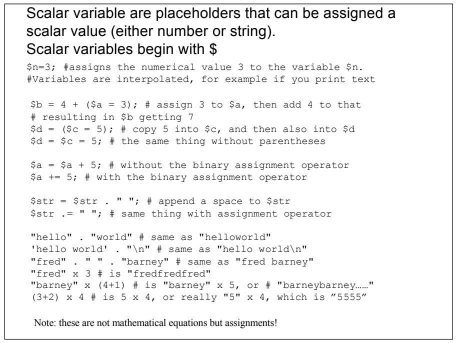


PERL conventions and rules

Basic Perl Punctuation:
line ends with ";"
empty lines in program are ignored
comments start with #
first line points to path to interpreter:

```
#! /usr/bin/perl
# "#!" is known as "shebang";
```

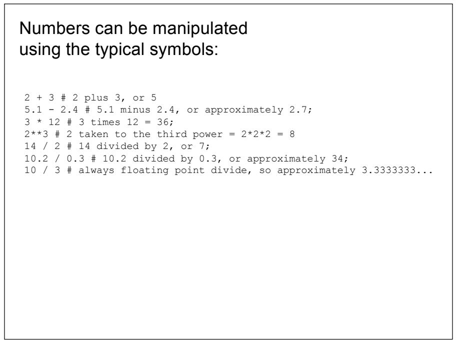
keep one command per line for readability
use indentation do show program blocks.
Variables start with $calars, @rrays, or %ashes
Scalars: foating point numbers, integers,
non decimal integers, strings



Scalar variable are placeholders that can be assigned a scalar value (either number or string). Scalar variables begin with $

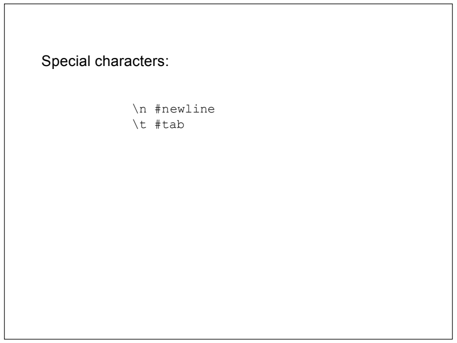
```
$n=3; #assigns the numerical value 3 to the variable $n.
#Variables are interpolated, for example if you print text

 $b = 4 + ($a = 3); # assign 3 to $a, then add 4 to that
 # resulting in $b getting 7
 $d = ($c = 5); # copy 5 into $c, and then also into $d
 $d = $c = 5; # the same thing without parentheses

 $a = $a + 5; # without the binary assignment operator
 $a += 5; # with the binary assignment operator

 $str = $str . " "; # append a space to $str
 $str .= " "; # same thing with assignment operator

 "hello" . "world" # same as "helloworld"
 'hello world' . "\n" # same as "hello world\n"
 "fred" . " " . "barney" # same as "fred barney"
 "fred" x 3 # is "fredfredfred"
 "barney" x (4+1) # is "barney" x 5, or # "barneybarney……"
 (3+2) x 4 # is 5 x 4, or really "5" x 4, which is ”5555”
```

Note: these are not mathematical equations but assignments!



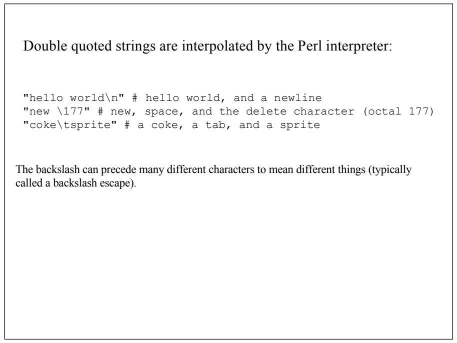
Numbers can be manipulated using the typical symbols:

```
2 + 3 # 2 plus 3, or 5
5.1 - 2.4 # 5.1 minus 2.4, or approximately 2.7;
3 * 12 # 3 times 12 = 36;
2**3 # 2 taken to the third power = 2*2*2 = 8
14 / 2 # 14 divided by 2, or 7;
10.2 / 0.3 # 10.2 divided by 0.3, or approximately 34;
10 / 3 # always floating point divide, so approximately 3.3333333...
```



Special characters:

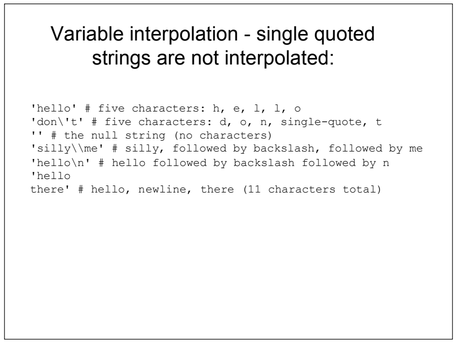
```
\n #newline
\t #tab
```



Double quoted strings are interpolated by the Perl interpreter:

```
"hello world\n" # hello world, and a newline
"new \177" # new, space, and the delete character (octal 177)
"coke\tsprite" # a coke, a tab, and a sprite
```

The backslash can precede many different characters to mean different things (typically called a backslash escape).



Variable interpolation - single quoted strings are not interpolated:

```
'hello' # five characters: h, e, l, l, o
'don\'t' # five characters: d, o, n, single-quote, t
'' # the null string (no characters)
'silly\\me' # silly, followed by backslash, followed by me
'hello\n' # hello followed by backslash followed by n
'hello
there' # hello, newline, there (11 characters total)
```