

## Bio 595

### Computers in Biomedical Research Fall 2003 Slides for Wednesday, December 3

## Hashes

Initialize a hash with key-value pairs- like initializing an array, but every pair becomes a key-value:

```
%classification = (  
  'dog',      'mammal',  
  'pigeon',   'bird',  
  'lizard',   'reptile',  
);
```

An equivalent way, showing pairing more clearly:

```
%classification = (  
  'dog'       => 'mammal',  
  'pigeon',   => 'bird',  
  'lizard'    => 'reptile',  
);
```

## Hashes

You can get an array of all the keys of a hash:

```
@keys = keys %my_hash;
```

You can also get an array of all the values of a hash:

```
@values = values %my_hash;
```

Hashes are a convenient data structure for accessing genetic data. Take *Homo sapiens*, with about 30,000 genes. Perhaps a DNA microarray experiment has been performed and has provided gene expression data for all these genes; for each gene (by name) you have an expression value. Unexpressed genes have the value 0.

Using an array, you could store only names of expressed (non-zero expression level) genes.

## Binary search

If, say, 8,000 genes are expressed, a lookup will require on the average looking at 4,000 gene names. If a gene name refers to a gene not expressed, the search will give a negative result. So, better to include all 30,000 genes and their expression levels. Searches will then average 15,000 lookups – slow.

An alternative: sorted arrays and binary search:

If gene names are arranged alphabetically, pick the middle element, discard the half depending on whether gene name is > or < the middle value; do this repeatedly until gene is found (or is not in the array)

Algorithm can use this to pick midpoint of the array:

```
$array[scalar[@array]/2]
```

## Compare and sort operations

To compare two strings alphabetically in Perl, use 'cmp':  
This operator returns 0 if strings are the same, -1 if they are in alphabetical order, +1 if in reverse order:

```
'ZZZ' cmp 'ZZZ'  returns 0  
'AAA' cmp 'ZZZ'  returns -1  
'ZZZ' cmp 'AAA'  returns +1
```

The binary search algorithm reduces the search to about 15 loops (instead of 15,000). Requires that the list be sorted to begin with. How do you sort an array?

```
@array = sort @array;
```

Sorting an array of numbers in ascending order:

```
@array = sort { $a <=> $b } @array;
```

## Gene expression data using hashes

Use hashes to find a gene in your data.

Load hashes so keys are the gene names and values are the expression levels.

A call to the hash with the name of the desired gene returns the expression level value.

Answer is returned much more quickly than from a binary search.

You can tell if the gene is in the data by saying

```
if( defined $myhash{'mykey'} ) {.....
```

If warnings are turned on, you'll get an error message (reference to an undefined value).

Unlike binary searches, a hash allows adding or subtracting elements without resorting the entire array.

## Hashes vs. arrays

Note that hashes don't store their elements in a sorted order. You can explicitly sort keys, however, like this:

```
@sorted_keys = sort keys %my_hash;
```

Use a hash if you need to see if something is in a set but don't need to list the set in order.

A sorted array plus the binary search algorithm is good if you need an ordered set and don't need routinely to add or subtract items often.

An array, used with 'push' and 'pop' functions, is a good way to get at the most recently added element(s).

An array, used with 'push' and 'shift', is good if elements don't need sorting but you need to add elements.

## Relational databases

Database programs: some are expensive (e.g. FileMaker Pro 6.0), but the one we'll use is free.

RDMS: Relational Database Management Systems – data are entered and extracted using SQL (Structured Query Language) to access data tables and their links.

Perl has a simple built-in mechanism for storing hash data: DBM, database management; it ties a hash to a file.

The task of translation:

Look at example ex25: translating triplets to amino acids. There are 3 versions, with different subroutine algorithms to match the triplet to the amino acid.

## Algorithms for translation

Note in these examples, the STDERR filename for saving errors (could also go to STDOUT to appear on screen)

Algorithm 1: returns amino acid one-letter code. Triplets are tested one at a time, all 64 possibilities.

Algorithm 2: account for redundancy and reduce the number of tests in the subroutine. Redundant codes usually have the same first 2 bases and vary in the 3<sup>rd</sup>. Note here that the amino acid list has been ordered.

Algorithm 3: use a hash; for each triplet key the amino acid is returned. Note the

```
if(exists $genetic_code{$codon})
```

returns 'true' if key \$codon exists, otherwise prints error.

## Comments on ex25v1-v3

Note how program loops through the DNA:

```
For(my $i=0 $i < (length($dna) - 2); $i += 3) {
```

Note that the 'for' loop has three parts delimited by semicolons.

First, the counter is initialized: my \$i=0 scopes this variable so it's visible only within the subroutine

Second, stop two from the end (you're reading triplets).

Quit reading if there aren't enough bases to read 3.

Third, increment by three.

The code

```
$codon = substr ($dna, $i 3);
```

extracts the 3-base triplet from the DNA at position \$i of length 3.

## DNA data file formats

As bioinformatics has developed and matured as a discipline, various data formats have been created and used: most important among the 20 or so such formats are GenBank and FASTA.

GenBank (Genetic Sequence Data Bank): a collection of all publicly released genetic data. Includes much information in addition to the sequence data. The most popular format.

FASTA and BLAST (Basic Local Alignment Search Technique) both use the FASTA format; the second most popular format.

EMBL (European Molecular Biology Laboratory): mostly the same data as GenBank, but slightly different format.

## DNA data file formats

DDB (DNA Data Bank of Japan) mostly same data as GenBank.

ABI (Applied Biosystems sequencer output): raw, unformatted DNA data hot off the machine.

PIR (Protein Identification Resource) protein sequence data.

GCG (Genetics Computer Group) from Accelrys Corp., earlier, widely used at universities with the GCG "Wisconsin" package; a different format.

GenBank and FASTA are the most common.

See the FASTA file called sample.dna.