

Bioinformatics: Life Sciences on Your Computer

Amino Acid Codes, Homology

This is supplemental reading to the video on Pairwise Sequence Alignment.

One-letter nucleotide codes

We should be familiar with the standard G, A, T and C for the four nucleotides in DNA. Those are the important ones. The following table includes ambiguity codes, used to either represent SNPs or to represent uncertainty in the sequencing. No need to memorize--here they are. The ambiguity codes are used far less frequently.

Symbol	Meaning	Explanation
G	G	Guanine
A	A	Adenine
T	T	Thymine (Uracil)
C	C	Cytosine
R	A or G	puRine
Y	C or T	pYrimidine
M	A or C	aMino
K	G or T	Keto
S	C or G	Strong interaction (3 bonds)
W	A or T	Weak interaction (2 bonds)
H	A, C or T (not G)	H follows G in alphabet
B	C, G or T (not A)	B follows A in alphabet
V	A, C or G (not T)	V follows TU in alphabet
D	A, G or T (not C)	D follows C in alphabet
N	A, C, G or T	aNything

One-letter amino acid codes

Here are the protein codes. In DNA, N is anything; in proteins, N is asparagine and X is anything. There are two rare amino acids--selenocysteine and pyrrolysine. There is also a Z code for glutamate/glutamine, which can be difficult to resolve. I won't quiz you on it, but it can be helpful to get to know the one-letter codes.

1-letter	Three-Letter Codes	Amino Acid
A	Ala	Alanine
C	Cys	Cysteine
D	Asp	Aspartic Acid
E	Glu	Glutamic Acid
F	Phe	Phenylalanine
G	Gly	Glycine
H	His	Histidine
I	Ile	Isoleucine
K	Lys	Lysine
L	Leu	Leucine

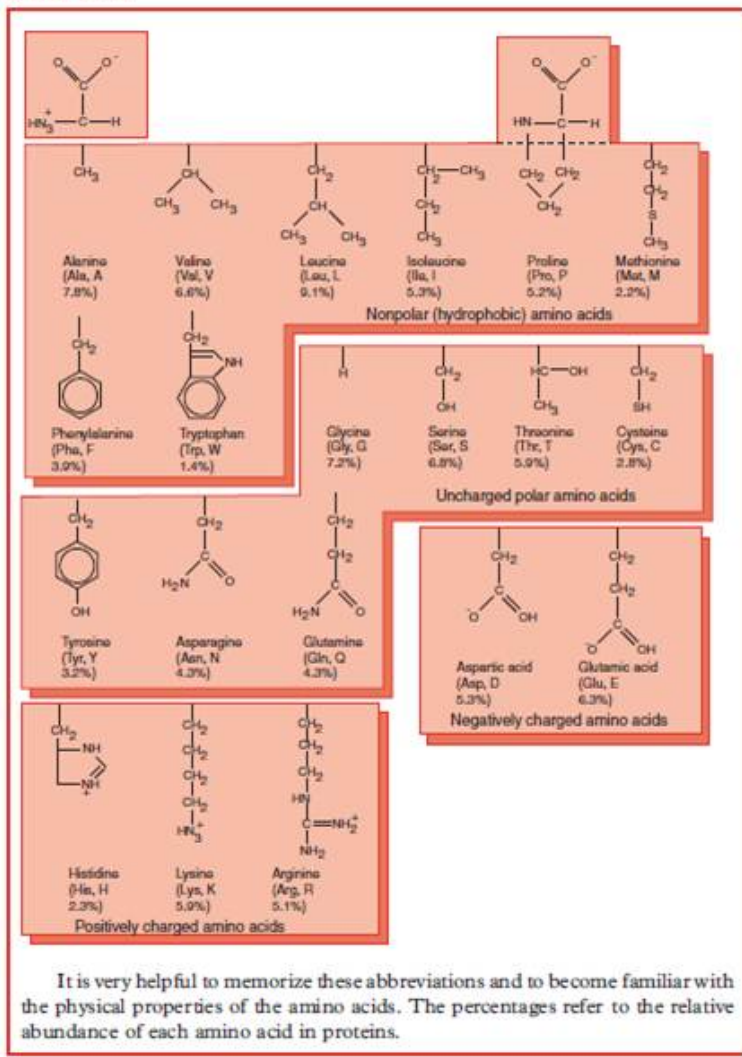
M	Met	Methionine
N	Asn	Asparagine
O	Pyl	Pyrrolysine
P	Pro	Proline
Q	Gln	Glutamine
R	Arg	Arginine
S	Ser	Serine
T	Thr	Threonine
U	Sec	Selenocysteine
V	Val	Valine
W	Trp	Tryptophan
X	Xaa	Undetermined
Y	Tyr	Tyrosine
Z	Gln	Glutamic acid or glutamine

Amino acids

Amino acids can be grouped by chemical properties, but there are several ways to do it and with some amino acids, the lines are a little blurred. The diagram below also shows the abundance of each amino acid in nature--note that leucine, alanine and glycine are very common while methionine and tryptophan are pretty rare. The structures are shown along with the one-letter codes. Selenocysteine and pyrrolysine are not shown.

Box 3.1

Structures and One- and Three-Letter Abbreviations of Twenty Common Amino Acids



Pairwise alignment

Pairwise sequence alignment is the optimal alignment of two DNA sequences or two protein sequences. It is a fundamental bioinformatics application, used to help determine if two proteins (or genes) are related structurally or functionally. It can be used to identify shared domains or motifs in a sequence or to find conserved regions of proteins.

Pairwise alignment is the basis of BLAST searching. BLAST is the Basic Local Alignment Search Tool--local alignment is part of the name. BLAST is used to start with a sequence and to find similar and related sequences in very large databases. To determine relatedness, it is necessary to be able to properly align the sequences in order to find the similar regions.

It is also used for the analysis of genomes--regions that are similar in two genomes tend to be coding regions of genomic DNA.

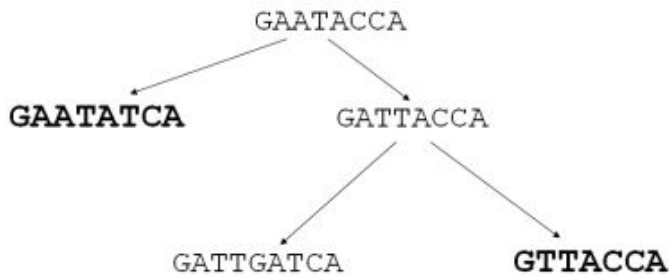
The goals of the alignments are important. An evolutionary study may focus on amino acid changes that occur frequently to find sequences that are close on evolutionary scale. A functional study may be more accepting of more rare amino acid changes to functionally similar amino acids.

Evolutionary distance

Let's look at the following sequences:

GAATATCA
GTTACCA

How far apart are those sequences? That depends on the ancestor sequence. If that can be deduced, an evolutionary distance can be calculated.



With a plausible ancestral sequence that can be determined by computer, you can see that one sequence can occur with one mutation, where it takes two steps, a substitution and a deletion to reach the other. That's an evolutionary distance of about three steps between the two original sequences.

Homology

During evolution, sequences diverge due to random mutation. **Similarity** is a quantitative measure used to help identify common ancestry. Changes to amino acids with functional roles tend to be preserved in nature. Sequences that are not critical to protein structure or gene regulation tend to be more prone to mutation, randomization. Patterns of conservation and variation can easily be identified. Because conserved regions tend to be less prone to mutation, conserved regions align better.

Sequences with a common ancestor are said to be **homologous**. Homology can not be quantified.

Definitions

homology: sequences share evolutionary history

similarity: includes non-exact matches

identity: measure of exact matches

orthology: homology in absence of gene duplication evidence (usually between species)

paralogy: homology resulting from gene duplication (usually within species)

analogous: sequences giving similar structural features

xenology: refers to sequence similarity caused by horizontal (inter-species) gene transfer

Discussion topic: Any confusion about the different homology definitions?

Similarity

Similarity refers to the extent to which nucleotide or protein sequences are related. The definition of **conservation** is crucial. When regions of two protein sequences have identical amino acids OR amino acids that are chemically similar, those regions are said to be conserved. A conservative substitution is a change from one amino acid to another amino acid in which physico-chemical properties are preserved. For instance, a change from arginine to lysine maintains the positive charge.

Identity refers to the fraction or more commonly the percentage of amino acids in a sequence that are identical. It is the extent to which two sequences are invariant--the amino acid (nucleotide) must be an exact match. Conservative substitutions can be quantified as well. Mathematically, similarity is identity plus conservation.

Example: Sequence A has 150 amino acids and Sequence B has 200 amino acids. An alignment shows that 45 amino acids are identical in the alignment.

$$\text{Identity} = 45 / 150 = 30\%$$

Always use the smaller sequence in the denominator.

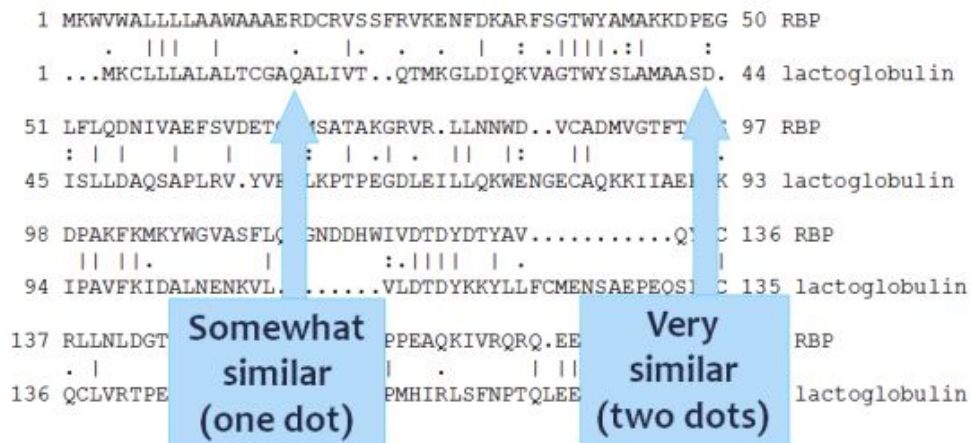
In addition to the 45 positions that are identical, there are 18 more where the amino acids are different but they are conservative substitutions that maintain chemical properties.

$$\text{Similarity} = (45 + 18) / 150 = 63 / 150 = 42\%$$

The sequences are said to be 30% identical and 42% similar. The similarity measure cannot be lower than the identity measure because similarity includes identity.

Output

The diagram below shows identity as vertical lines. Two levels of similarity are shown as well. Not all alignment programs use the same notation.

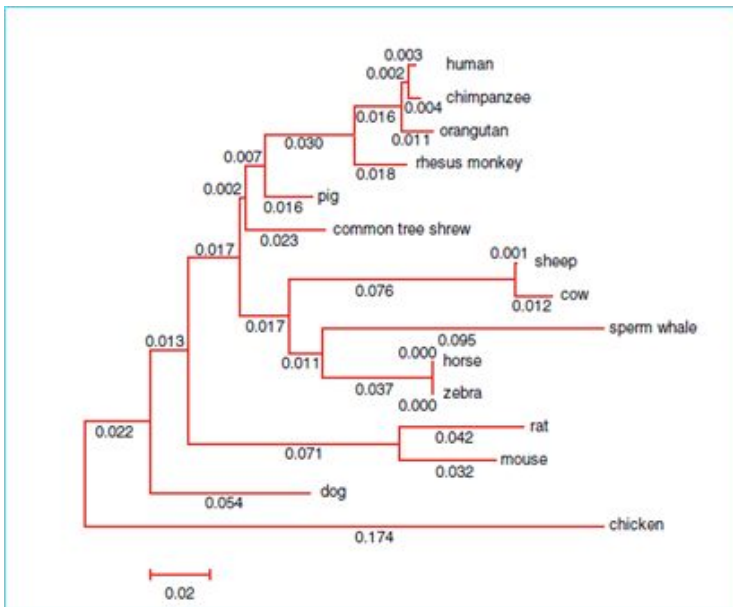


Different programs have different output notations

Orthologs

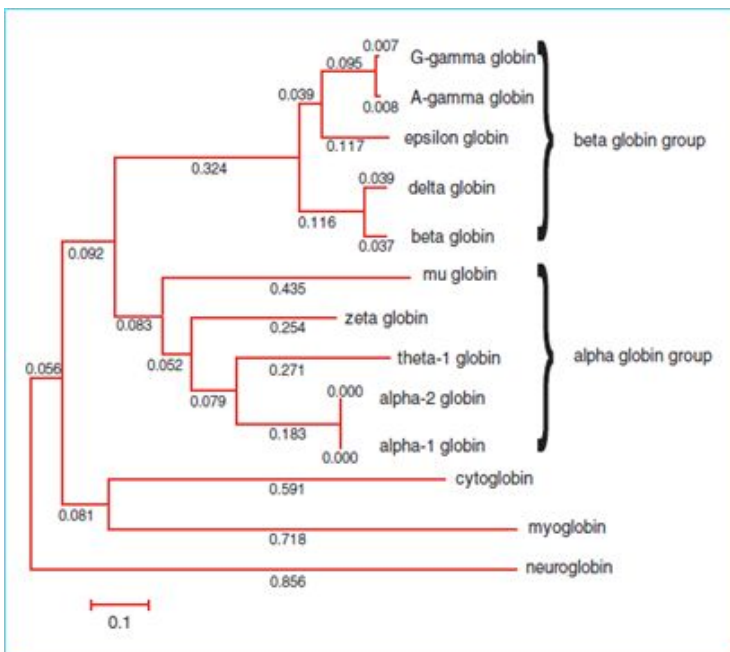
The diagram below shows homologs from different species. Each arose from a common gene. Speciation separated genes. For instance, all mammals have a gene for beta globin. We know that beta globin has been around a long, long time. Each mammal has a beta globin because that protein was around in the very first mammals. As species evolve, all needed that beta globin. The sequences, however, changed a bit through random mutation as species evolved. So human beta globin is very similar to horse beta globin, but human beta globin is even more similar to chimpanzee beta globin. The common ancestor of chimp and human is much more recent than the common ancestor of human and horse.

Because orthologs usually developed their function before speciation, orthologs are presumed to have similar function.



Paralogs

Paralogs are the result of gene duplication. The term is usually reserved for genes/proteins of the same species. However, the paralogs diverged a long, long time ago. When a gene duplicates, often over time, the genes evolve to take on two separate functions. Human globins are shown below.



Closer relative example

So which are usually the closer relatives: orthologs or paralogs?

Example

Suppose that long before mammals existed, there were five important genes. We will call them :

A, B, C, D, E

Think of them as crayons of different colors. Pick your colors but A is red.

```
Query: 181 catcaactacaactccaagacacccttacaccactaggatatcaacaaacctaccac 240
          ||||| ||| ||||| |||| | ||||| ||||| ||||| ||||| |||||
Sbjct: 189 catcaactgcaaccccaaagccaccct-caccactaggatatcaacaaacctaccac 247
```