Comparative Genomics I

1. Background
Some definitions:

Homologs – general term for genes/proteins that have similar sequence and are derived from a common ancestral sequence

Orthologs – homologs derived through speciation

Paralogs – homologs derived through gene duplication

The above may sometimes be written as "homologues", "orthologues", and "paralogues" in British English, or one may refer to "homologous proteins" or "orthologous genes", *etc*.

Example: human alpha-globin and chimpanzee alpha-globin are orthologs. Human alpha-globin and human beta-globin are paralogs. Human alpha-globin is more similar in sequence to chimpanzee alpha-globin than it is to human beta-globin – gene duplication predates speciation.

Analogs – genes with similar sequence due to convergent evolution (not common ancestry). This is very rare at the sequence level, but there can be functional analogs, such as proteins that have non-homologous sequences but perform the same molecular function.

Homology search – matching a given sequence to other known genes or proteins in a database. This is typically the first step after a genome sequence has been determined and the genes predicted. Now that many complete genomes have been sequenced, it is common to compare one genome to another and determine how many genes they have in common. For distantly related species, it is often impossible to distinguish orthologs and paralogs.

2. BLAST – Basic Local Alignment Search Tool.
The most commonly-used homology search tool.
Does not align complete sequences, but finds subsequences with the best possible alignment.

For protein sequences:
identical: amino acids are the same
positive: amino acids may be different, but have similar biochemical properties (size, charge)

Matches are scored by "E-value", where E represents the number of matches expected at random when searching a database of a given size. E = 1 means 1 match would be expected at random from the database. This is similar to a probability (P-value). $E = 10^{-6}$ means that there is only a 1/million chance of observing such a match at random. The lower the E value, the greater the confidence that the sequences are homologous

For quick searches of the genome, some genome browsers, such as UCSC (http://genome-test.cse.ucsc.edu/), use BLAT (BLAST-like alignment tool). It is similar, but not the same as BLAST. BLAT uses a faster algorithm based on 11-mers (=11 bases of DNA) or 4-mers (=4 amino acids) to find matches of:
95% or greater identity over 25 bases or more (DNA)
80% or greater identity over 20 amino acids or more (Proteins)

A number of alignment tools have been developed to very quickly map "short reads" (typically DNA sequences in the range of 35-250 bases) to a reference genome. These are used for next generation sequencing data. Typically, they align complete sequences and expect a very close match with the reference sequence. Examples: BWA, bowtie, Stampy, NextGenMap

Molecular evolutionists/systematists are usually interested in comparing orthologs. However, it is difficult to distinguish orthologs and paralogs from homology searches. One method that is often used is "reciprocal best hits". This approach compares genes in two genomes using two steps:

i) Gene A from species 1 is used for a BLAST search of the species 2 genome. The best match (or "hit") is gene A'

ii) Gene A' from species 2 is then used for a BLAST search of the species 1 genome. If the best match is gene A, then these are reciprocal best hits and are considered orthologs.

Note that the above method does not guarantee that genes A and A' are true orthologs, the result could be misleading if there are independent gene duplication/loss events in the two species.

Another approach is to consider only "one-to-one" orthologs. These are defined as homologous genes that occur in only one single copy in each genome.

3. Prokaryotic Comparative Genomics

Prokaryotic comparative genomics is much more advanced than eukaryotic comparative genomics. Over 40,000 complete prokaryotic genomes are publicly available, and this number is rapidly increasing. Almost all have been sequenced by WGS (whole-genome shotgun sequencing). Reasons:
a) small genome (0.6 - 10 Mb, single circular chromosome)
b) little repetitive DNA (easier assembly)
c) no introns (easier gene prediction)
d) medical/economic importance (e.g., human pathogens)

For up-to-date numbers and information, see:
http://bacteria.ensembl.org/index.html

Prokaryotes are divided into two major domains (or empires) of life, which allows very diverse comparisons:
a) Bacteria (or Eubacteria) - common commensal and pathogenic bacteria
b) Archaea (or Archaebacteria) - ancient group of mostly extremophiles (live at high temp, *etc*.)

4. Gene Loss in Pathogenic Bacteria

Example: *Mycobacterium leprae* (leprosy) vs. *M. tuberculosis*(tuberculosis*)*

M. leprae appears to have lost the function of half of its genes. This may explain why it has the longest doubling time of any known bacteria and why it cannot be cultured in the laboratory.

pseudogenes – previously protein-encoding genes that have mutations that disrupt the ORF (insertion of stop codon or insertion/deletion that causes frameshift)

In general, pathogenic *Mycoplasma* species have the smallest genomes known in free-living bacteria, as low as 580 kb. However, this represents an evolutionary derived state. That is, they are not primitive bacteria that have gained only the genes necessary for survival. Their ancestors had many more genes that were lost over the course of *Mycoplasma* evolution.

Other, unrelated pathogens also have small genomes:
Rickettsia sp.- Rocky Mountain spotted fever, Mediterranean spotted fever in humans (1.2 Mb)
Borrelia burgdorferi - causes Lyme disease in humans (1.4 Mb)

also the symbiotic bacteria of aphids, *Buchnera aphidicola* have a very small genome.

Obligate association with host tissues promotes genome reduction.

Which genes are lost? Many genes involved in energy metabolism are lost. Metabolic intermediates and energy sources are taken from the host. Many genes required for amino acid and vitamin synthesis are lost. These are also provided by the host. An exception in the aphid symbiont, *B. aphidicola*, 10% of its genes are involved in synthesis of essential amino acids not synthesized by the host. However, it has lost the genes needed to synthesize amino acids that are made by the host. Thus, it is a true symbiont.

The smallest endosymbiont bacterial genome known is from the genus *Carsonella* (160 Kb, 182 genes). These bacteria live in sap-eating insects, which have a low-protein diet. Over half of the genes in the *Carsonella* genome are involved in translation and amino acid metabolism.

Why are genes lost? Two possibilities:
a) Selective advantage for smallness? Bacteria with smaller genomes can replicate faster and outcompete those with larger genomes that replicate slower. This does not appear to be the case:
- small changes in DNA content (gene-sized) do not appear to affect replication rate
- many pathogens retain non-functional pseudogene DNA (e.g., *M. leprae*)
- small genomes are not more densely packed than large (same amount of "spacer DNA")

b) Mutation pressure
The major reason for gene loss is thought to be a mutational. If there is not selective pressure to maintain a gene, it will eventually be lost due to mutation.
- there is a bias towards deletions
- there is a bias towards AT mutations

5. Hyperthermophile Comparative Genomics
hyperthermophiles – live at 80-100 C, mostly Archaebacteria, some Eubacteria
thermophiles – live at 50-65 C, mostly Archaebacteria, some Eubacteria
mesophiles – under 50 C, mostly Eubacteria, some Archaebacteria

Are there specific genes that allow survival at very high temperatures? That is, are there genes specific to hyperthermophiles?

To test this, one can search the COG (Cluster of Orthologous Group) database:
http://www.ncbi.nlm.nih.gov/COG/

for proteins present in hyperthermophiles, but not in thermophiles or mesophiles. Importantly, one must consider evolutionary relationships in the analysis. For example, there are some Archaea that

do not live at high temperatures and some Eubacteria that do live at high temperature. In other words, the presence of the gene should follow the temperature, not the phylogenetic relationship.

The result: one protein out of 2791 is specific to hyperthermophiles:

Reverse gyrase is a large protein (>1000 a.a.) that is a fusion of two protein domains, helicase and topoisomerase. It introduces twists into double-stranded circular DNA and may help prevent unwinding of DNA at high temperatures.