Sequence Similarity Searches on the World Wide Web
Stuart M. Brown, NYU Medical Center, New York, NY, USA
Internet: browns02@mcrcr6.med.nyu.edu

If you have just determined the sequence of an interesting bit of DNA, one of the first questions you are likely to ask yourself is “has anybody else seen anything like this before?” Fortunately, there has been a very successful international effort to collect all known sequences (both DNA and protein) into databases so that they can be searched. These databases (and searching tools) are freely available on the Web. The problem is that these databases are HUGE and, as a result, you must compare your sequence with the vast number of other sequences.

Sequence comparison is the most powerful and most reliable method of answering biological questions about the evolutionary relationships between genes. Recent improvements in the statistical methods used by similarity searching computer programs allow the biologist to make conclusions with a high level of certainty. Now that the entire genomes of some organisms have been completely sequenced, it is possible to use similarity-search methods to provide definitive answers to questions such as “Is there a copy of this interesting gene in this organism?”

A database search is a similarity search, but it is frequently—although incorrectly—referred to as homology searching. The term “homology” implies a common evolutionary relationship between two traits. Just because two sequences share a stretch of nearly identical nucleotides (or amino acids [aa]) does not mean that they are directly descended from a common ancestor. Homologous proteins share a common 3-dimensional structure, while proteins with a chance similarity of aa sequence do not. Over evolutionary time, two homologous proteins may diverge to the point that sequence similarity cannot be detected, yet they may still retain their common structure. Homologous proteins generally have similar biological functions, but this is not a rigorous requirement; conversely, discovery of homology is not proof of function.

Of course, a high level of similarity is a strong indication of homology. As a rule of thumb, 25% identity over a stretch of 100 aa can be considered to be good evidence of common ancestry for two sequences. Homologous sequences are usually similar over an entire sequence, or sometimes just over one functional domain, but it is never correct to state that two sequences are “50% homologous.” Homology is an all or nothing decision. Be careful in asserting homology between short regions of two sequences; matches that are more than 50% identical in a 20–40 aa region can occur by chance.

At present, the two most popular programs for similarity searching are Basic Local Alignment Search Tool (BLAST) (1) and FASTA (3). They use similar approaches to reach similar answers; although, there are some subtle differences. The two programs will generally give slightly different results: in terms of which sequences are found to be similar, their relative rankings by similarity score and the statistical significance that is assigned to those similarity scores. It is generally thought that FASTA is more sensitive for DNA-DNA comparisons, but BLAST is usually faster. The choice of which program to use is often governed by factors of convenience rather than mathematical rigor; but to be thorough, it is best to perform searches with both programs.

For most researchers, the Web has become the best method to make similarity searches. Web servers are available to everyone, they have the most current data, they are usually fast and your computer is not tied up while your search is computing. However, the addresses and the services offered by various Web servers (See BioBit) are subject to change without warning, so it is wise to have some alternatives lined up when your computing is dependent on the kindness of strangers.

Protein similarity searches can find much more distant similarities than comparisons of DNA sequences (ca. 2.5 billion vs. 100 million years of evolutionary divergence). This is true for several

news.group.news

News.group.news goes to cyberspace and back to gather Net news for you. We emphasize practical, methods-related issues that are discussed in many newsgroups and also feature issues of general interest to biological scientists. But be warned! On this page of BioTechniques, we don’t review it — we just report it!

Experiencing difficulty cloning a CDNA insert in E. coli, presumably due to toxicity, a netter on bionet.molbio. methods-reagnts was advised to clone the insert in lytic bacteriophage lambda. Netters highly recommend the protocol for direct sequencing from lambda that is available on the Web site of The Institute for Genomic Research, www.tigr.org

Seeking to extract mRNA from single Arabidopsis leaves, a netter was advised to wrap individual leaves in aluminum foil, drop the foil packets into liquid nitrogen and pulverize the tissue by tapping packets with a hammer. Tissue pulverized in this manner can be readily transferred to a chilled pestle containing ice-cold extraction buffer.

Seeking information on restriction enzymes, a netter was advised to visit REBASE at www.neb.com/rebase, a comprehensive database of information about restriction enzymes and their associated methylases, including recognition and cleavage sites and commercial availability.

Inquiring about the effect of primer mismatches on Taq processivity, a netter was advised to consult R. Sommer and D. Tautz, 1989, Minimal homology requirements for PCR primers, Nucleic Acids Res. 17:6749.

BioTechniques Editorial Department
reasons. (i) The DNA “alphabet” is restricted to only 4 letters, while the aa alphabet has 20 letters, so the probability of chance matches is much greater with DNA-DNA comparisons. (ii) Two differing DNA bases can only be scored as a mismatch, while two aas can share varying degrees of similarity based on their physical and chemical properties etc. (iii) The protein databanks are much smaller than the DNA databanks, so searches can be more sensitive without incurring too many false positives. Generally, if your query sequence is protein, you will search protein databanks, and with DNA sequence, you will search nucleotide data. However, it is possible to automatically translate your DNA sequence into aa in all six reading frames (BLASTX and FASTX) and compare it to protein databases, or to compare your protein sequence to the six reading frame translations of all DNA database sequences (TFASTA and TBLASTN). Surprisingly, a protein-protein search using a translated DNA sequence as a query against a protein databank, or a protein query against a translated DNA databank, is much more sensitive than a DNA-DNA search. It is generally better to perform all searches as protein-protein comparisons unless you are studying DNA sequences that do not code for protein. For some searches, it is important to find every possible database match to a given sequence; for others, it may be advantageous to limit the search to just humans, to mammals or to the animal kingdom. The measurement of statistical significance in similarity searches is dependent on the size of the database. As the database gets bigger, weak homologies become less significant (i.e., the chance of finding equally strong matches due to random similarities increases). The more you can restrict your search, the faster it will run, the better your significance scores will be and the fewer false-positive hits you will have to sort through. The scores reported by similarity programs are called “expectation values”, but they should be treated just like the “P values” from traditional types of statistical tests. These “expectation values” represent the number of times that a match with the observed similarity score would occur by chance if the database actually contained no sequences that were truly similar to your query sequence (i.e., in a database of random sequences). Generally, expectation values less than 0.05 or 0.02 are considered significant. The results of a similarity search can be divided into three categories: (i) highly significant matches between nearly identical proteins, (ii) insignificant matches that are clearly due to chance similarities and (iii) an intermediate range of matches with scores at the borderline of statistical significance. These intermediate values have been described as the “twilight zone” (2). Each “twilight zone” protein can be used as the query for another similarity search. True homologues will have significant matches, not only with your original query sequence but also to other members of that protein family (i.e., other sequences that received highly significant scores in the original search). Unrelated sequences will not match any other proteins in the family of your original query sequence (4). BLAST and FASTA produce similar output files. First there is a short description of the program and a list of the databases and program options chosen. Then there is a list of all of the database sequences that matched your query sequence. A number is assigned to each of these sequences that represents the quality of the match. The list is presented in descending order, so that the best matches are at the top of the list. However, the most biologically significant matches are not always the ones ranked highest in the list. A quick look at this list of hits can provide a lot of information.

**REFERENCES**