An Introduction to Sequence Similarity ("Homology") Searching

William R. Pearson[1]

[1]University of Virginia School of Medicine, Charlottesville, Virginia

ABSTRACT

Sequence similarity searching, typically with BLAST, is the most widely used and most reliable strategy for characterizing newly determined sequences. Sequence similarity searches can identify "homologous" proteins or genes by detecting excess similarity—statistically significant similarity that reflects common ancestry. This unit provides an overview of the inference of homology from significant similarity, and introduces other units in this chapter that provide more details on effective strategies for identifying homologs. *Curr. Protoc. Bioinform.* 42:3.1.1-3.1.8. © 2013 by John Wiley & Sons, Inc.

Keywords: sequence similarity • homology • orthlogy • paralogy •
sequence alignment • multiple alignment • sequence evolution

AN INTRODUCTION TO IDENTIFYING HOMOLOGOUS SEQUENCES

Sequence similarity searching to identify homologous sequences is one of the first, and most informative, steps in any analysis of newly determined sequences. Modern protein sequence databases are very comprehensive, so that more than 80% of metagenomic sequence samples typically share significant similarity with proteins in sequence databases. Widely used similarity searching programs, like BLAST (Altschul et al., 1997; *UNIT 3.3 & 3.4*), PSI-BLAST (Altschul et al., 1997), SSEARCH (Smith and Waterman, 1981; Pearson, 1991; *UNIT 3.10*), FASTA (Pearson and Lipman, 1988; *UNIT 3.9*), and the HMMER3 (Johnson et al., 2010) programs produce accurate statistical estimates, ensuring protein sequences that share significant similarity also have similar structures. Similarity searching is effective and reliable because sequences that share significant similarity can be inferred to be homologous; they share a common ancestor.

The units in this chapter present practical strategies for identifying homologous sequences in DNA and protein databases (*UNITS 3.3, 3.4, 3.5, 3.9, 3.10*); once homologs have been found, more accurate alignments can be built from multiple sequence alignments (*UNIT 3.7*), which can also form the basis for more sensitive searches, phenotype prediction, and evolutionary analysis.

While similarity searching is an effective and reliable strategy for identifying homologs—sequences that share a common evolutionary ancestor—most similarity searches seek to answer a much more challenging question: "Is there a related sequence with a similar function?" The inference of functional similarity from homology is more difficult, both because functional similarity is more difficult to quantify and because the relationship between homology (structure) and function is complex. This introduction first discusses how homology is inferred from significant similarity, and how those inferences can be confirmed, and then considers strategies that connect homology to more accurate functional prediction.

INFERRING HOMOLOGY FROM SIMILARITY

The concept of homology—common evolutionary ancestry—is central to computational analyses of protein and DNA sequences, but the link between similarity and homology is

Current Protocols in Bioinformatics 3.1.1-3.1.8, June 2013
Published online June 2013 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/0471250953.bi0301s42

often misunderstood. We infer homology when two sequences or structures share more similarity than would be expected by chance; when excess similarity is observed, the simplest explanation for that excess is that the two sequences did not arise independently, they arose from a common ancestor. Common ancestry explains excess similarity (other explanations require similar structures to arise independently); thus, excess similarity implies common ancestry.

However, homologous sequences do not always share significant sequence similarity; there are thousands of homologous protein alignments that are not significant, but are clearly homologous based on statistically significant structural similarity or strong sequence similarity to an intermediate sequence. Thus, when a similarity search finds a statistically significant match, we can confidently infer that the two sequences are homologous, but if no statistically significant match is found in a database, we cannot be certain that no homologs are present. Sequence similarity search tools like BLAST, FASTA, and HMMER minimize false positives (nonhomologs with significant scores; Type I errors), but do not make claims about false negatives (homologs with nonsignificant scores; Type II errors). As is discussed below, it is often easier to detect distant homologs when searching a smaller ($<$100,000–500,000 entry) database than when searching the most comprehensive sequence sets (more than 10,000,000 protein entries). Likewise, when domain annotation databases like InterPro and Pfam annotate a domain on a protein, it is almost certainly there. But these databases can fail to annotate a domain that is present, because it is very distant from other known homologs.

We infer homology based on excess similarity; thus, statistical models must be used to estimate whether an alignment similarity score would be expected by chance. Today, comprehensive protein databases contain tens of millions of sequences, the vast majority of which are unrelated to an individual query. Thus, it is very easy to determine the distribution of scores expected by chance, and it has been observed that *unrelated* sequences have similarity scores that are indistinguishable from *random* sequence alignments. For local sequence alignments, like those produced by BLAST, Smith-Waterman, or FASTA, the expected distribution of similarity scores by chance (scores for alignments between two random or unrelated sequences) is described by the extreme value distribution $p(\mathrm{s} \geq x) \leq 1 - exp[-exp(-x)]$ (Fig. 3.1.1), where the score "*s*" has been normalized to correct for the scaling of the scoring matrix (*UNIT 3.5*) and the length of the sequences being compared. To avoid these normalization issues, most similarity searching programs also provide a score in bits, which can be converted in to a probability using the formula:

$$p(b \geq x) \leq 1 - \exp(-mn2^{-x})$$

where *m,n* are the lengths of the two sequences being aligned. For scores with $p()<0.01$, which will include any significant score in a database search, this expression can be simplified to $p(b \geq x) = (mn2^{-x})$. However, the probability $p(b)$ is not what is reported by BLAST, FASTA, or SSEARCH, because it reflects the probability of the score in a *single* pairwise alignment. Current search programs report the best scores after hundreds of thousands to tens of millions of comparisons have been done; as a result BLAST and other programs report the expected number of times the score would occur by chance—the e-value, E()-value, or expectation value—after thousands or millions of searches. The E()-value is $E(b) \leq p(b)D$, where D is the number of sequences in the database. BLAST actually uses a slightly different correction that has the same effect.

Because the expectation value depends on database size, an alignment score found by searching 10,000,000-entry database will be 100-fold less significant than exactly the same score found in a search of a 100,000 entry database. This does not mean that sequences can be homologous in one context (the smaller search) but not in another. If the alignment was significant in the smaller (and shorter) search, the sequences are

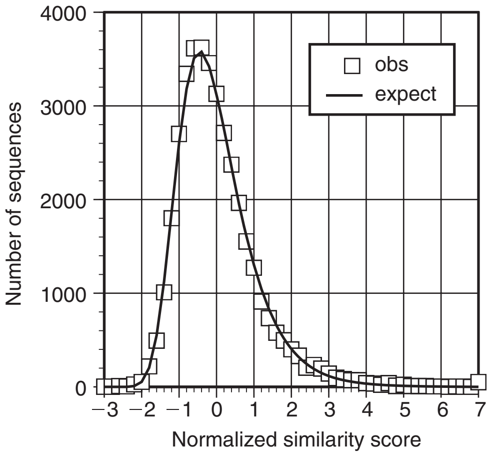

Figure 3.1.1 The distribution of real and expected similarity scores. The human dual specificity protein phosphatase 12 (DUS12_HUMAN) was compared to 38,114 human RefSeq proteins using the SSEARCH program. The distribution of bit-scores (or standard deviations above and below the mean 0) for all 38,114 alignments is shown (squares, □), as well as the mathematically expected distribution of *z*-scores based on the size of the database, using the extreme-value distribution. The close agreement between the observed and expected distribution of scores reflects the observation that the distribution of unrelated sequence scores is indistinguishable from random (mathematically generated) scores, so sequences with significant sequence similarity can be inferred to be not-unrelated, or homologous.

homologous, but that homology may not be detected in the larger search because there are 100-fold more sequences that could produce high (but not significant) alignment scores by chance. Sequences that share significant sequence similarity can be inferred to be homologous, but the absence of significant similarity (in a single search) does not imply nonhomology. For nonsignificant alignments, comparisons to an intermediate sequence, or analysis with profile or HMM-based methods, can be used to demonstrate homology.

The most common reason homologs are missed is because DNA sequences, rather than protein sequences (or translated DNA sequences), are compared. Protein (and translated-DNA) similarity searches are much more sensitive than DNA:DNA searches. DNA:DNA alignments have between 5- to 10-fold shorter evolutionary look-back time than protein:protein or translated DNA:protein alignments. DNA:DNA alignments rarely detect homology after more than 200 to 400 million years of divergence; protein:protein alignments routinely detect homology in sequences that last shared a common ancestor more than 2.5 billion years ago (e.g., humans to bacteria). Moreover, DNA:DNA alignment statistics are less accurate than protein:protein statistics; while protein:protein alignments with expectation values <0.001 can reliably be used to infer homology, DNA:DNA expectation values $<10^{-6}$ often occur by chance, and 10^{-10} is a more widely accepted threshold for homology based on DNA:DNA searches. The most effective way to improve search sensitivity with DNA sequences is to use translated-DNA:protein alignments, such as those produced by BLASTX and FASTX, rather than DNA:DNA alignments.

Reliable homology inferences require reliable statistical estimates. The statistical estimates provided by BLAST, FASTA, SSEARCH, and other widely used similarity searching programs are very reliable. But in unusual cases, which can appear scientifically exciting, the statistical estimation fails, and unrelated sequences are assigned statistical significance not because of homology, but because of statistical errors. When a scientifically unexpected alignment appears to be statistically significant, investigators should consider alternate strategies for estimating statistical significance. Statistical estimates can be confirmed in two ways: (1) by attempting to identify the most similar alignments to unrelated sequences and confirming that those alignments are not significant; or (2) running additional searches with shuffled versions of the original sequences.

Since it is difficult to be certain that two sequences are unrelated, particularly for unexpected inferences of homology, strategy (1) above might seem impractical. Fortunately, high-scoring alignments that happen by chance can be quite diverse, as they do not reflect evolutionary relationships. Alignments between unrelated sequences will have very different domain content and structural classifications. By examining the domain structures (and possibly structural classes) of high-scoring alignments, one can identify the highest scoring unrelated sequences because sequence alignments with proteins containing unrelated domains must be unrelated. Accurate statistical estimates will give E()-values $\sim$1 to unrelated sequences (sequences with different domains); if unrelated sequences have E()-values <0.001-0.01, then the scientifically novel relationship is suspect.

The second strategy—inferring significance from shuffled sequences with identical length and composition—is much more straightforward, but it depends on the assumption that shuffled sequences have similar properties to real protein sequences. This is certainly true for most sequences, but the assumption may be weaker in the exceptional cases that produce “novel” results. The most reliable shuffling strategies preserve local residue composition, either by simply reversing one of the sequences or by shuffling the sequences in local windows of 10 or 20 residues, which preserves local amino acid composition. SSEARCH (*UNIT 3.10*) and the other members of the FASTA program package (*UNIT 3.9*) offer statistical estimates based on shuffles that preserve local composition.

Homology (common ancestry and similar structure) can be reliably inferred from statistically significant similarity in a BLAST, FASTA, SSEARCH, or HMMER search, but to infer that two proteins are homologous does not guarantee that every part of one protein has a homolog in the other. BLAST, SSEARCH, FASTA, and HMMER calculate *local* sequence alignments; local alignments identify the most similar region between two sequences. For single domain proteins, the end of the alignment may coincide with the ends of the proteins, but for domains that are found in different sequence contexts in different proteins, the alignment should be limited to the homologous domain, since the domain homology is providing the sequence similarity captured in the score. When local alignments end within a protein, the ends of the alignment can depend on the scoring matrix used to calculate the score. In particular, scoring matrices like BLOSUM62, which is used by BLASTP, or BLOSUM50, which is used by SSEARCH and FASTA, are designed to detect very distant similarities, and have relatively low penalties for mismatched residues. As a result, a homologous region that is 50% identical or more can be extended outside the homologous domain into neighboring nonhomologous regions. This is a common cause of errors with iterative methods like PSI-BLAST (Gonzalez and Pearson, 2010), but can be reduced by limiting extension in later iterations (Li et al., 2012). The relationship between the similarity scoring matrix and alignment overextension is discussed in *UNIT 3.5*.

E()-values, identity, and bits

While homology is inferred from excess similarity, and excess similarity is recognized from statistical estimates [E()-values], most investigators are more comfortable describing similarity in terms of “percent identity”. Although a common rule of thumb is that two sequences are homologous if they are more than 30% identical over their entire lengths (much higher identities are seen by chance in short alignments), the 30% criterion misses many easily detected homologs. While 30% identical alignments over more than 100 residues are almost always statistically significant, many homologs are readily found with E()-values $<10^{-10}$ that are not 30% identical. Thus, E()-values and bit-scores (see below) are much more useful for inferring homology. When 6,629 *S. cerevisiae* proteins were compared to 20,241 human proteins, 3,084 of the yeast proteins shared significant [E()$<10^{-6}$] similarity with a human protein, but only 2,081 of those proteins were more than 30% identical. There are 19 yeast proteins whose closest human homolog shares less than 20% identity [E()-values from 10^{-7}–10^{-40}; about an equal number are $>$80% identical]. A 30% identity threshold for homology underestimates the number of homologs detected by sequence similarity between humans and yeast by 33% (this is a minimum estimate; even more homologs can be detected by more sensitive comparison methods).

The bit-score provides a better rule-of-thumb for inferring homology. For average length proteins, a bit score of 50 is almost always significant. A bit score of 40 is only significant [E() $<$ 0.001] in searches of protein databases with fewer than 7000 entries. Increasing the score by 10 bits increases the significance 2^{10}=1000-fold, so 50 bits would be significant in a database with less than 7 million entries (10 times SwissProt, and within a factor of 3 of the largest protein databases). Thus, the NCBI Blast Web site uses a color code of blue for alignment with scores between 40 to 50 bits, and green for scores between 50 to 80 bits. In the yeast versus human example, the alignments with less than 20% identity had scores ranging from 55 to 170 bits. Except for very long proteins and very large databases, 50 bits of similarity score will always be statistically significant and is a much better rule-of-thumb for inferring homology in protein alignments.

While percent identity is not a very sensitive or reliable measure of sequence similarity– E()-values or bits are far more useful—percent identity is a reasonable proxy for evolutionary distance, *once homology has been established*. Like raw similarity scores, bit-scores and E()-values reflect the evolutionary distance of the two aligned sequences, the length of the sequences, and the scoring matrix used for the alignment (*UNIT 3.5*). An alignment that is twice as long, e.g., 200 residues instead of 100 residues at the same evolutionary distance, will have a bit score that is twice as high. Since the E()-value is proportional to 2^{-bits}, a two-fold higher bit-score squares the E()-value (10^{-20} becomes 10^{-40}). For analyses that depend on evolutionary distance, percent identity provides a useful approximation, but evolutionary distance is not linear with percent identity. The evolutionary distance associated with a 10% change in percent identity is much greater at longer distances. Thus, a change from 80% to 70% identity might reflect divergence 200 million years earlier in time, but the change from 30% to 20% might correspond to a billion year divergence time change.

INFERRING FUNCTION FROM HOMOLOGY

Homologous sequences have similar structures, and frequently, they have similar functions as well. But the relationship between homology and function is less predictable, both because a single protein structure can have multiple functions (does chymotrypsin have a different “function” from trypsin?) and because the concept of “function” is more ambiguous (the same enzyme have “different” roles in two tissues because of different concentrations of substrates).

Currently, the most popular strategy for inferring functional similarity is to focus on “orthologs,” typically understood as the same protein in different organisms. The concept of “orthology” was originally introduced to distinguish two kinds of evolutionary histories: (1) homologous sequences that differ because of speciation events—orthologs; and (2) homologous sequences that were produced by gene-duplication events—paralogs (Koonin, 2005). Orthologs were initially recognized in phylogenetic reconstructions because they could accurately reproduce the evolutionary histories of the organisms where they were found. Because paralogs are duplicates that preserve the original gene (presumably with the original function), paralogous proteins are more likely to acquire novel functions. But "novelty" very much depends on the level of functional specificity. For example, trypsin and chymotrypsin are paralogous serine proteases with slightly different substrate specificities, but many of their Gene Ontology terms do not capture their functional differences; thus, they are paralogs that are functionally very similar.

Unfortunately, the terms "orthologous" and "paralogous" have been given different meanings in different contexts. Their original phylogenetic definitions—orthology, same gene, different organism; paralogy, duplicated genes—did not have clear functional implications. More recently, orthology has been associated with functional similarity (sometimes in the absence of homology, e.g., functional orthologs), while paralogy has sometimes been defined as functionally distinct (Gerlt and Babbitt, 2000). Perhaps because of the greater interest in homologs with similar functions, the term orthologous has been applied more generously than evolutionary analyses might support, and paralogous genes are assumed to have different functions. From the evolutionary perspective, both orthologs and paralogs are details about evolutionary history that have some ability to improve function prediction, but for many protein families most paralogs have similar functions.

The problem of functional inference is exacerbated by the need to infer function at great evolutionary distances; the average matches between mouse and human proteins share about 84% identity, but yeast-human closest matches are about 30% identical on average. It is easy to argue that two proteins that are more than 80% identical with conserved active sites will share the same function, but more difficult to reliably infer similar function at much longer evolutionary distances. For example, humans have three trypsin paralogs to TRY1_HUMAN: TRY2_HUMAN and TRY6_HUMAN, which are both more than 90% identical, and TRY3_HUMAN, which is about 70% identical. In addition, there are about 50 other paralogs that range from 45% to 25% identical. CTRB2_HUMAN is the most similar chymotrypsin, at 38% identity. All these proteins share significant global similarity and are identical at the three residues in the serine protease catalytic triad; thus, we can infer that these paralogs share the serine protease function. Several studies have shown that homologous sequences that share more than 40% identity are very likely to share functional similarity as judged by E.C. (Enzyme Commission) numbers, but counter-examples exist where a small number of residues in very similar proteins are associated with dramatic changes in enzyme activity. Inferring functional similarity based solely on significant local similarity is less reliable than inferences based on global similarity and conserved active site residues.

FROM PAIRWISE TO MULTIPLE SEQUENCE ALIGNMENT

Pairwise sequence alignments, such as those calculated by BLAST, FASTA, and SSEARCH, view the evolutionary structure of a protein or domain family from a single perspective. Pairwise alignments produce very accurate statistical significance estimates, so one can have great confidence when significant homologs are found. But the one-sequence perspective has shortcomings as well; searches with models of protein families, using either PSI-BLAST or Hidden Markov Model (HMM)–based methods, can identify far more homologs in a single search at little additional computational cost. Moreover, the multiple sequence alignments that are used to construct the position-specific scoring matrices of PSI-BLAST or the Hidden Markov Models used by HMMER provide important information about the most conserved regions in the protein; locating conserved regions in protein and domain families can dramatically improve predictions of the functional consequences of mutation. Multiple sequence alignments thus provide much more structural, functional, and phylogenetic information than pairwise alignments.

While multiple sequence alignments are much more informative, they cannot be used to answer the first critical question about two sequences—are they homologous? The inclusion of a protein into a multiple sequence alignment requires independent evidence for homology; multiple sequence alignment programs do not provide statistical estimates and will readily align nonhomologous sequences (particularly nonhomologous sequences that are highly ranked by chance in a similarity search). The assumption of homology can be especially misleading with iterative methods like PSI-BLAST, because once a nonhomologous domain has been included in the multiple sequence alignment used to produce the position-specific scoring matrix, the matrix can become re-purposed towards finding members of the nonhomologous family. Ideally, each sequence included in a multiple sequence alignment will be evaluated both to ensure that it shares significant similarity with some of the other members of the family, and that the boundaries of the included sequence correspond with the boundaries of the domain homology.

Rigorously building a multiple sequence alignment is exponentially more computationally expensive than pairwise alignment. Rigorous pairwise alignment algorithms require time proportional to the product of two sequences [$O(n^2)$], so increasing the sequence lengths 2-fold increases the time required 4-fold. Fortunately, protein sequences have a limited range of lengths, so rigorous searches (SSEARCH) are routine. In contrast, the rigorous multiple sequence alignment of 10 sequences of length 400 would take proportional to 400^{10} and 100 sequences would take $O(400^{100})$, so rigorous multiple sequence alignment is impractical. During the 1980s, progressive alignment strategies, like ClustalW (Larkin et al., 2007; *UNIT 2.3*) were developed that simplified the problem to $O(n^2l^2)$, where n is the number of sequences, and l is their average length. These early progressive alignment strategies suffered from the problem that gaps placed early on in the alignment could not be re-adjusted to reflect information from sequences aligned later. More recent multiple sequence alignment methods, like MAFFT (Katoh et al., 2002) and MUSCLE (Edgar, 2004), use iterative approaches that allow gaps to be re-positioned.

There is a much greater diversity of multiple sequence alignment algorithms than pairwise sequence alignment algorithms, largely because optimal pairwise solutions are readily available, but multiple sequence alignment strategies use different heuristic approximations. While different multiple sequence alignment programs will often produce modestly different results, most programs produce very similar results for sequences at modest evolutionary distances (greater than 40% identity), and the differences are found near the boundaries of gaps. A more common problem is multiple sequence alignments with large gaps, which may reflect the presence or absence of domains in a subset of the sequence set. Alignment only makes biological sense when the residues included in the alignment are homologous. Large differences in sequence length, or attempts to multiply align sequences that are locally, but not globally, homologous can produce very different results because the programs are aligning nonhomologous domains.

SUMMARY

BLAST, FASTA, SSEARCH, and other commonly used similarity searching programs produce accurate statistical estimates that can be used to reliably infer homology. Searches with protein sequences (BLASTP, FASTP, SSEARCH,) or translated DNA sequences (BLASTX, FASTX) are preferred because they are 5- to 10-fold more sensitive than DNA:DNA sequence comparison. The 30% identity rule-of-thumb is too conservative; statistically significant [$E() < 10^{-6} - 10^{-3}$] protein homologs can share less than 20% identity. E()-values and bit scores (bits >50) are far more sensitive and reliable than percent identity for inferring homology.

With the rise of whole-genome sequencing, protein sequence databases are both comprehensive and large. It is rarely necessary to search complete sequence databases to find close-homologs; more sensitive searches can be limited to complete protein sets from evolutionarily close organisms. Because of its sensitivity and the slower changes in protein sequences, protein similarity searching can easily detect vertebrate homologs, almost all homologous sequences that have diverged in the past 500 million years (bilateria), and a very large fraction of the sequences that diverged in the past billion years. The most efficient and sensitive searches will focus on well-annotated model organisms sharing a common ancestor that diverged in the past 500 million years.

Because of the abundance of relatively closely related (>40% identical) sequences in comprehensive databases, the accuracy and location of annotation can often be more important than finding the closest homolog. The SwissProt subset of the UniProt database is unique in providing comprehensive information on modified residues, active sites, variation, and mutation studies that allow more accurate functional prediction from homologous alignments. Once an homologous protein or domain has been found, establishing the state of functionally critical residues (and ensuring that functional domains are part of the alignment) can greatly decrease errors produced by simply copying the name (and function) of the reference protein to the query sequence. Similarity searches that use statistical significance to well-curated sequences provide the most accurate functional predictions.

LITERATURE CITED

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* 25:3389-3402.

Edgar, R.C. 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32:1792-1797.

Gerlt, J.A. and Babbitt, P.C. 2000. Can sequence determine function? *Genome Biol.* 1:REVIEWS0005.

Gonzalez, M.W. and Pearson, W.R. 2010. Homologous over-extension: A challenge for iterative similarity searches. *Nucleic Acids Res.* 38:2177-2189.

Johnson, L.S., Eddy, S.R., and Portugaly, E. 2010. Hidden markov model speed heuristic and iterative hmm search procedure. *BMC Bioinformatics* 11:431.

Katoh, K., Misawa, K., Kuma, K., and Miyata, T. 2002. MAFFT: A novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic Acids Res.* 30:3059-3066.

Koonin, E.V. 2005. Orthologs, paralogs, and evolutionary genomics. *Ann. Rev. Genet.* 39:309-338.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J., and Higgins, D.G. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics* 23:2947-2948.

Li, W., McWilliam, H., Goujon, M., Cowley, A., Lopez, R., and Pearson, W.R. 2012. PSI-Search: Iterative HOE-reduced profile Ssearch searching. *Bioinformatics* 28:1650-1651.

Pearson, W.R. 1991. Searching protein sequence libraries: Comparison of the sensitivity and selectivity of the smith-waterman and FASTA algorithms. *Genomics* 11:635-650.

Pearson, W.R. and Lipman, D.J. 1988. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. U.S.A.* 85:2444-2448.

Smith, T.F. and Waterman, M.S. 1981. Identification of common molecular subsequences. *J. Mol. Biol.* 147:195-197.