Technical Report

Database Searches with Multiple Oligopeptides Containing Ambiguous Residues

BioTechniques 21:1116-1117 (December 1996)

Michael H. Vodkin, Robert J. Novak and Gerald L. McLaughlin[1]

Illinois Natural History Survey, Champaign and University of Illinois, Urbana, IL;[1]Indiana University School of Medicine, Indianapolis, IN, USA

Several techniques in molecular biology frequently yield partial and ambiguous data on genes and gene products. For instance, N-terminal sequence analysis of oligopeptide cleavage products generates this type of sequence data. Typically, data generated from blotted or HPLC-resolved peptides consist of disconnected and unordered oligopeptides derived from N-terminal analysis of fragments resulting from complete or partial trypsin, chymotrypsin or CNBr digestion; such sequences are also "linked" if they were derived from the same isolated polypeptide, e.g., a band identified after sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) and blotting. To empirically identify the protein represented by such data, labor-intensive sequencing, with or without cloning, is frequently required. We became interested in defining a strategy to more reliably identify the source protein from existing sequence databases without an investment of additional laboratory experiments.

Several algorithms are readily available to rapidly search the databases for proteins or nucleic acids that are identical to or related to a specified query sequence. One popular program is the basic local alignment search tool (BLAST) (Reference 1; see Availability). However, BLAST can search databases (e.g., SWISS-PROT) with only moderate sensitivity. At the National Institutes of Health (NIH) address (see Availability), BLAST can search the updated, nonredundant protein or nucleic acid databases.

A disadvantage of BLAST is that very limited ambiguity is allowed at each position. For amino acids, "X" designates an unknown, "B" designates aspartate or asparagine, "Z" designates glutamate or glutamine and "-" designates a gap of indeterminate length. Table 1 shows an actual example of such data. When the individual oligopeptides listed in Table 1 were used to search the SWISS-PROT database, multiple related and unrelated sequences with similar or identical scores were retrieved. Even by comparing the individual lists for common, multiple hits, it was not possible to determine a unique candidate protein that was related to all or most of the oligopeptides.

Another approach tested was to search with BLAST in pairwise or N-wise combinations of the oligopeptides, either as a continuous string of residues or as a broken string with hyphens designated as a discontinuity. (BLAST at the NIH supports the latter syntax; however, BLAST at some other addresses does not.) The first method created strings of characters that were not originally juxtaposed and thus did not allow the correct identification. The second method, when used in various pairwise combinations, still did not detect homologous proteins in the database.

We therefore utilized an alternative search algorithm and show its utility for identifying a protein in the database when query sequences include several linked oligopeptide fragments with some ambiguous amino acid residues. FindPatterns, or Find (a subset of the GCG package; see Availability), was used for the peptide data set in Table 1. Find has more versatility for managing ambiguous residues and multiple, discontinuous oligopeptides. Each individual residue of the query oligopeptide can be specified as either unknown (X) or up to a 20-fold ambiguity (every amino acid candidate at a position is encoded, as in Table 1). The gap size between the unordered fragments can also be specified with a minimum

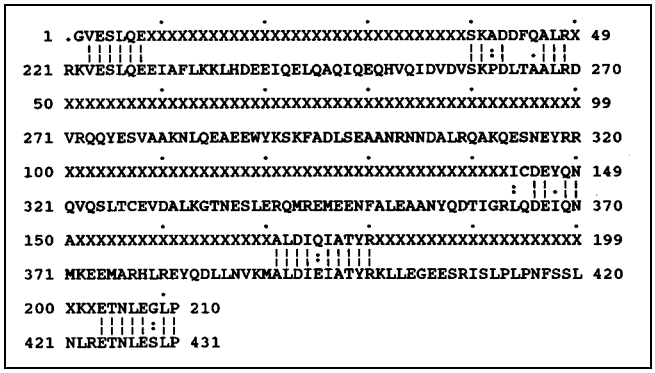

Figure 1. Alignment of ordered oligopeptides to rat vimentin. The oligopeptides listed in Table 1 were searched in various pairwise combinations against the SWISS-PROT database with FindPatterns, Mismatch=4, minimal spacing between oligopeptides set at two and maximal set at 500 residue (GCG package). The following is an actual example of the syntax, using oligopeptides 1 and 2, in both orientations as the query (after $ or other prompt): find/batch/mis=4/pat=(K,R)X(E,L)TNLEGLP(X){2,500}I(C,R)DEYQ(N,K)A find/batch/mis=4/pat=I(C,R)DEYQ(N,K)A(X){2,500}(K,R)X (E,L)TNLEGLP. After it was determined that all the oligopeptides except for the tripeptide number 5 could be aligned with the available vimentin sequences, a protein was artificially created to maintain the appropriate spacing as determined by each oligopeptide's matched position, using X's for unknown amino acids, and to resolve ambiguous amino acid residues maximizing identity between the query and rat vimentin sequence (SW:VIME_RAT). This artificial protein (top line) was then aligned to rat vimentin (bottom) with GAP (GCG package).

Table 1. N-terminal Sequence of Tryptic Peptides of a Protein from Rat-Derived *P. carinii*

Oligopeptide No.	Residues
1	(K,R)X(E,L)TNLEGLP
2	I(C,R)DEYQ(N,K)A
3	(Q,A,Y)(E,L)(D,S)(L,I)QIATYR
4	(T,S,L)(A,K)(A,F)(D,S)(D,T)FQALR
5	QFP
6	(G,N,S)(V,L)(Q,E,Y)(D,S)L(Q,E)(E,I)

A protein was isolated using detergent extraction and column chromatography from the lungs of rats heavily infected with *P. carinii*. The protein was resolved on a SDS-polyacrylamide gel and found to have a mobility corresponding to 40–60 kDa. It was blotted to an Immobilon® filter (Millipore, Bedford, MA, USA), eluted and partially cleaved with trypsin. Each fragment was resolved by HPLC, and N-terminal analysis was performed on 6 fragments. Each amino acid residue is designated by a single letter. Multiple letters within parentheses indicate ambiguity at that site in the biochemical analysis. "X" is an unknown residue.

and maximum number of residues. The maximum, in the case of chemical or proteolytic digestion products of an unknown but purified peptide or protein, was set as slightly greater than the molecular weight of the protein as determined by relative mobility on SDS-PAGE (in this example of a 40–60-kDa polypeptide set at 500 residues).

Three of the 10 tested pairwise combinations, with all ambiguities inserted, jointly detected the family of eukaryotic vimentins (intermediate filament proteins) in one of the two possible orientations of the oligopeptides. Indeed, all of the oligopeptides listed on Table 1, except for the tripeptide 5, were successfully aligned and oriented with the vimentin sequences of vertebrates when searched with the identified vimentin proteins; Figure 1 shows the example of rat vimentin. Nineteen of the 28 unambiguous residues (scored as one of 3 for oligopeptide 5), or 68%, were identical to the aligned rat vimentin. For 3 of the 5 (60%) aligned oligopeptides, a potential R or K cleavage site was also located 1 or 2 residues upstream of the N-terminal sequence, as predicted by the specificity of trypsin (2). Taken together, these data indicate that the protein isolated from the infected rat lung is definitely related to vimentin. However, the amount of divergence for the unambiguous residues suggests that it is not identical to bona fide rat or to any other published vertebrate vimentin. No sequences from invertebrates are currently available for comparison; we hypothesize that it represents the *Pneumocystis carinii* vimentin analog, but another possibility is a more distant member of the rat vimentin family.

The probability of encountering multiple oligopeptides in an individual database protein is the product of the oligopeptides' individual probabilities. One reason for our success of matching the query peptides to the database with Find was the ability to incorporate more of the data, i.e., to also include the ambiguous residues in the search. Another reason is that the pairwise searching algorithm practically eliminates the chance of encountering a random single, very similar short peptide sequence in the same database. In our experience, this type of problem has increased significantly as the sequence databases have grown in size, and new approaches such as those we have used are necessary to address such random false homologies.

The strategy that we have defined to address the problem with discontinuous oligopeptides can be refined and extended to oligonucleotide sequences. For instance, similar searches can be designed to determine whether short oligonucleotides (e.g., random-amplified polymorphic DNA [RAPD] primers) are linked in the same nucleotide database entry and to help identify origins of linked but discontinuous DNA sequences. In summary, especially when multiple, short, ambiguous and linked sequences are available, the use of more versatile algorithms, such as Find with pairs of sequences, can provide a powerful advantage for database searching.

AVAILABILITY

BLAST is accessible from Genetics Computer Group (GCG), Program Manual for the Wisconsin Package, Version 8.1, July 1995, 575 Science Dr., Madison, WI 53711, USA, E-mail (**blast@ncbi.nlm.nih.gov**) or at several World Wide Web sites (the primary one for this study is at the BCM Search Launcher, Human Genome Center, Houston, TX: **http://gc.bcm.tmc.edu:8088/search-launcher/launcher.html**).

ACKNOWLEDGMENTS

The primary peptide sequence data came from Dr. S.F. Queener and Dr. Lori Bolyard, Department of Pharmacology, Indiana University School of Medicine, Indianapolis, IN, and we thank them for permission to use the data to demonstrate the search strategy. We also wish to thank Dr. J. Siegel, Illinois Natural History Survey and Dr. R. Weigel, University of Illinois, Department of Veterinary Pathobiology for reading earlier drafts of the manuscript. This work was supported in part by a grant to R.J.N. from the Illinois Department of Natural Resources (SENR TM34). Computer support was provided to M.H.V. from the Pittsburgh Supercomputer Center (PSCB DMB890077P).

REFERENCES

1. **Altschul, S.F., W. Gish, W. Miller, E.W. Myers and D.J. Lipman.** 1990. Basic local alignment search tool. J. Mol. Biol. *215*:403-410.
2. **Voet, D. and J. Voet.** 1995. Biochemistry, 2nd ed. John Wiley & Sons, New York.

Received 12 March 1996; accepted 10 May 1996.

Address correspondence to:
Gerald McLaughlin
Indiana University School of Medicine
Department of Pathology and Laboratory Medicine
MS A128
Indianapolis, IN 46202-5120, USA.
Internet: gmclaugh@indyvax.iupui.edu