IDENTIFYING UNKNOWN PROTEINS

Winona C. Barker and Margaret O. Dayhoff

National Biomedical Research Foundation
Georgetown University Medical Center
Washington, D.C. 20007

Abstract

In this paper we discuss ways to identify a protein, both when its amino acid sequence is known and, particularly, prior to the determination of the complete sequence. If a similar sequence is in the Protein Sequence Database, an unknown may be identified on the basis of partial or ambiguous sequence data, or on the basis of amino acid composition. Identification in the early stages of structural determination can save time and scarce resources by preventing duplicate effort or by suggesting optimal sequencing strategies based on similar known sequences. Moreover, for these methods to become clinically useful, identification must be made on the basis of minimal data.

Introduction

The human body contains hundreds of cell types that need to be examined with regard to the identity and quantity of proteins that they produce in normal and disease states. The characterization of these many proteins and their variants will be expedited by a combination of special experimental design and computer analysis. The accompanying paper describes some of the technical aspects of an on-line protein identification system utilizing an extensive knowledge base of protein structures and ancillary information and a battery of analytical and predictive computer programs. Many of the components of this system are well developed and are being used every day. This paper briefly describes the applications of these methods to identify sequenced proteins, as well as those where the available experimental data are partial or ambiguous sequences or amino acid compositions.

The problem of identifying an unknown protein is one that many biomedical researchers frequently face. Two aspects to identification are assigning a meaningful name to the protein and determining if this protein is the same as, a mutant form of, or closely related to another known protein. Classical protein chemists assign names to proteins on the basis of function, location, or other properties that form at least part of the basis for isolating and purifying the proteins. Names describing the functions of proteins often do not uniquely specify the protein: there are many ribonuclease, some related and some apparently quite unrelated.

Examples of names based partly on the tissue, organ, or organism from which the protein is isolated include platelet factor 4, flagellin, 30S ribosomal protein, and botulinus toxin. Some names based on other properties are proline-rich phosphoprotein, basic blue protein, and allergen H1. Among the least informative protein names are substance P, protein C, and J chain. As proteins are better characterized, their original common names may be replaced by more specific designations. Old Yellow Enzyme is now recognized as NADPH dehydrogenase, myokinase is adenylate kinase, and rhodanese is thiosulfate sulfurtransferase.

A number of proteins at first isolated from specific tissues or organs have later been found to be identical with or very similar to proteins found elsewhere. Vasoactive intestinal peptide, a hormone secreted by the intestinal mucosa, is one of a number of gastrointestinal hormones that now are known or suspected to also function as neurotransmitters. Urogastrone, isolated from human urine, is epidermal growth factor. The nuclear protein A24, found in all chromatin examined, is a dimer of ubiquitin and histone H2A.

The advent of techniques to rapidly sequence nucleic acids has led to the inference of many amino acid sequences for which the name, function, or even the actual existence of a corresponding protein is not known. These sequences have been called open reading frames, unidentified reading frames, or hypothetical proteins.

Identifying a New Sequence

We have classified all of the known protein sequences into superfamilies of related sequences. Within a superfamil, sequences are
grouped into families that are less that 50% different, subfamilies that are less that 20% different, and entries that are less that 5% different. The first step in deciding if a new sequence belongs in a known superfamily or represents a new superfamily is to compare selected segments, usually about 25 residues long and including the amino end and any known active sites, with all segments of the sequences currently in our Protein Sequence Database using the program SEARCH. For sequence comparisons we commonly use a scoring matrix that we have derived from the frequencies with which pairs of amino acids were observed to replace each other in closely related proteins.

If in the search a particular protein is retrieved with a high score for a number of the segments, an alignment of the two sequences often provides convincing evidence of their relatedness. This is usually possible if 40% of the amino acids are identical over a stretch of 25 or more residues, without introducing gaps in either sequence. To evaluate the possibility of more distant relationship, we use two computer programs, RELATE and ALIGN. In the first, all segments of a given length in one sequence are compared with all segments in the second sequence. In the other, the best alignment of two segments is made. For both methods, a scoring matrix is required and a numerical property of the comparison is calculated. This same property is also calculated for a large number of pairs of permuted sequences (with the same amino acid compositions as the real sequences). The mean and standard deviation of the property are estimated from the distribution of scores of the permuted sequences. A real score occurring by chance can then be made for comparisons as the real sequences). The mean and standard deviation of the property are estimated by Doolittle.

Our criterion for judging that two sequences belong in the same superfamily is that the probability of the similarity occurring by chance is less that $10^{-6}$, which corresponds to a score of nearly 5 SD when using RELATE or ALIGN. This rather low probability has been chosen so that in making all possible comparisons of 2000 sequences, each representing one group of clearly related sequences, the number placed into the wrong superfamily would be very small, approximately four. When there is additional evidence, such as similarity of active site, function, or three-dimensional structure, sequences may be placed in the same superfamily on the basis of scores above 3 SD, corresponding to a probability of $10^{-3}$. These criteria for defining a superfamily of proteins provide an objective method for establishing protein relationships. Nevertheless, some proteins classified into different superfamilies, such as the V and C regions of immunoglobulins, have no doubt had a very distant common evolutionary origin. Many of the problems involved in the statistical detection of very distant relationships have recently been discussed by Doolittle.
Identification Using Partial Sequences

There are certain experimental protocols that yield partial sequences. Very commonly the amino ends of an intact protein chain and of its cyanogen bromide fragments will be subjected to sequenator analysis, producing sequences of segments 20-40 residues long. A search of the Protein Sequence Database will easily identify the sequence if it is in the database. Moreover, a sequence that is within the same family (less than 50% different) would almost always be noticed in a search of 5 segments. Inasmuch as sequences would have to change very rapidly in order for homologous proteins in different mammals to be more than 50% different, the presence of a homolog from another mammal in the database would usually ensure the identification of the human sequence.

To illustrate, we chose eight of the most rapidly changing proteins for which we have both human and other mammalian sequences. Non-overlapping 25-residue segments of these were searched against a database from which all of the human sequences had been removed, leaving 1905 sequences. In all, 42 searches were performed. The highest scores for homologous segments ranged from 59 to 149, whereas the highest scores for segments from unrelated proteins ranged from 35 to 56.

The probability of finding a closely related protein in the database depends on many factors, including the probability of success in searching for a single segment, the number of segments actually searched, and the fraction of possible types of proteins currently represented in the database. We have discussed these probabilities in detail elsewhere. We conjecture that, at present, between 25% and 50% of all the protein superfamilies to be found in free-living organisms are represented in the database. For some major taxonomic groups, however, the number of families and superfamilies of proteins represented is quite small.

Identification Using Ambiguous Sequence Data

Occasionally the initial attempt to subject a newly isolated protein to sequenator analysis will yield two amino acids at each step because the sample is a mixture of two proteins or of two chains, each with a free amino end. In principle, the two components of such a mixture can be easily identified if they are both already in the Protein Sequence Database. For instance, a mixture of prealbumin and alpha-antitrypsin would give the following information for each position:

```
1 2 3 4 5 6 7 8 9 10 11 12
Gly Asp Thr Gly Gly Asp Ala Ala Lys Cys Pro Asp
Glu Pro Pro Gin Thr Gly Glu Ser Gin Lys Thr Leu
```

The database could be searched for all sequences that contain Gly or Glu followed by Asp or Pro followed by Thr or Pro, etc. If 12 positions are used, both prealbumin and alpha-antitrypsin would match perfectly and no other sequence currently in the database would match at more than 8 positions.

In most cases, identification of only 6-8 amino acid residues (not necessarily contiguous) will uniquely identify a sequence if it is in the database. Such an identification made in the early stages of a sequence determination would save time and money by preventing redundant efforts. Thus, microsequencing of peptides containing certain radiolabeled amino acids also yields sufficient information to identify a protein. We were asked to identify a partially sequenced protein isolated from pituitary tissue. In the first 33 residues, Lys, Arg, or Met had been identified at eight of the positions. Nothing was known about the sequence at other positions except that these three amino acids did not occur. Only one sequence, ubiquitin, matched the known information perfectly. We have suggested a new strategy to optimize the probability of identifying radiolabeled proteins. A tissue is incubated with a mixture of all necessary amino acids under conditions that permit protein synthesis. Several of the amino acids carry one radioactive label, several carry another label, and the rest are unlabeled. Proteins synthesized by the tissue can be separated and the label carried at each of the first 30 or so positions in the sequence can be determined. With such data, most human proteins could be identified if the sequences were in the database.

Identification on the Basis of Amino Acid Composition

It is surprisingly easy to recognize a protein in the collection knowing only the molecular weight and the amino acid composition. For example, suppose we can estimate the amount of each amino acid within 20% or 2 residues, whichever is larger.

Using the known values for human cytochrome c, we searched successively for sequences with matching values, with the following results:

```
<table>
<thead>
<tr>
<th>Number of Sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entire database</td>
</tr>
<tr>
<td>% Lys 13.8 to 20.8</td>
</tr>
<tr>
<td>% Gly 10.0 to 15.0</td>
</tr>
<tr>
<td>% Ser 0 to 3.9</td>
</tr>
<tr>
<td>% Ile 5.7 to 9.7</td>
</tr>
</tbody>
</table>
```

Similarly, for rabbit actin, a sequence of rather average composition:

```
<table>
<thead>
<tr>
<th>Number of Sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entire database</td>
</tr>
<tr>
<td>Mol. Wt. 33452-50179</td>
</tr>
<tr>
<td>% Lys 4.0 to 6.2</td>
</tr>
<tr>
<td>% Gly 6.0 to 9.0</td>
</tr>
<tr>
<td>% Ser 4.8 to 7.4</td>
</tr>
<tr>
<td>% Ile 6.4 to 9.6</td>
</tr>
<tr>
<td>% Arg 3.8 to 5.8</td>
</tr>
</tbody>
</table>
```

Preliminary tests indicate that, if the amino acid composition (percentage of each amino acid)
is known exactly for either the new sequence or the comparison data (as when it is derived from the sequences), these data will be very powerful. It is frequently possible to measure the percentage composition of the amino acids in a protein to a precision of 2%. Usually the composition includes 18 values, aspartic acid being combined with asparagine and glutamic acid with glutamine. The probability is less than 4 x 10^{-10} of finding a similar composition by chance. Because only 100,000 proteins are to be found in the whole human genome, such a similarity would be quite sufficient to identify a sequence by similarity of its composition.

Before an amino acid sequence is determined, the amino acid composition is usually analyzed. Many of these data already in the literature have been reviewed and tables compiled by Kirchenbaum (see, for example, reference 16). These data may contain a considerable amount of experimental error and may even eventually be found to be mixtures of proteins. We are investigating the usefulness of these preliminary data for identification of newly examined proteins. We are setting up a database of protein compositional data in a retrieval system so that we can readily compare these data with compositions derived from protein sequences.

Frequently the composition of a small active peptide is determined. As this peptide may have been derived by chemical degradation from a much larger sequence, no structure with this composition will be found as a separate entry. In such a case, all of the sequences must be searched for segments with the proper amino acid composition. We plan to modify the SEARCH program to accomplish this.

Identification using amino acid composition is based on much more information than is a typical identification using an antibody. This latter method sometimes leads to errors. An example of identifying an unknown protein from amino acid composition data was reported by scientists in Roger Guillemin's laboratory at the Salk Institute. 17 A protein immunologically "identified" as the precursor of corticotropin (ACTH) and beta-endorphin was isolated and the amino acid compositions of the protein and several tryptic peptides were determined. A search of the known protein sequences clearly identified the protein as an IgG1 heavy chain. The unexpected immunological cross-reactivity was apparently the result of similarities in portions of the amino acid sequences of these two molecules.

Concluding Remarks

High resolution methods to separate hundreds of proteins obtained from a single cell type are currently being developed. For instance, plates from two-dimensional gel electrophoresis, following the method of O'Farrell, can now be reproduced reasonably well and the corresponding spots on different plates located by a calibration involving reference spots. 18-21 The method is attractive because the differences in two samples, as for example the production of an abnormal protein, can be visualized and the amount of each protein produced can be measured. For these methods to be useful in a clinical setting, it is important that identification of proteins be possible on the basis of minimal data as it will usually be impractical to subject the proteins from individual patients to complete sequence analysis.

This work was supported by NIGMS grant GM08710 and NICHD grant HD09547.

References

12. Barker, W.C., Dayhoff, M.O., Evolution of


