Better Cutters for Protein Mass Fingerprinting: Preliminary Findings

Michael J. Wise[1], Tim Littlejohn[2] and Ian Humphery-Smith[3]

1. European Bioinformatics Institute, Hinxton, UK (On leave from Department of Computer Science, University of Sydney, Australia, 2006).

2. Australian Genome Information Centre, University of Sydney, Australia 2006

3. Centre for Proteome Research and Gene-Product Mapping, Australian Technology Park, Eveleigh, Australia, 1430.

Keywords: proteome, peptide-mass fingerprinting, protease, computer search

Abstract

Peptide-Mass Fingerprinting (PMF) encompasses a number of techniques for protein characterization which have as a first step the cleaving of target proteins by chemical or enzymatic reagents. Software systems exist which perform similar analyses. However, this is the first study which examines theoretically the effectiveness of the particular reagents for PMF. In this study, the task of PMF was to identify every sequence in a non-redundant protein database, via "in silico" digestion with theoretical proteases. From these experiments, some conclusions are drawn about the characteristics of better reagents and the experimental conditions which are more likely to be useful for PMF. The need for strongly non-redundant databases is also highlighted.

1. Introduction

Peptide-Mass Fingerprinting (PMF) has rapidly found applications within molecular biology for the confirmation of recombinant protein expression and the detection of insertions, deletions and mutations within genes and gene products. There are software systems which perform similar analyses, e.g. PROPSEARCH (Hobohm et al., 1994). Input to this system are the percentage composition values for 16 amino acids (Asn and Asp are combined; Gln and Glu are combined; Trp and Cys are ignored). In addition, the system uses values for the isoelectric point and the total molecular weight of the unknown protein. The root-mean-square distance between the experimental values and the database values is then used to retrieve and order related proteins.

Another software system often used for protein identification is MOWSE (Pappin et al., 1993). The MOWSE system has shown how experimentally derived mass fragment data can be used to discover related proteins within databases by using 3 or 4 experimentally determined peptide masses as a query against a database derived from theoretical digestion with the same proteases. MOWSE is tolerant of slight variations in mass without loosing sensitivity, indicating that fragment mass is an extremely useful indicator of overall protein similarity.

However, to our knowledge, no study has appeared which examines theoretically the effectiveness of different proteases for protein-mass fingerprinting. That is, by doing experiments on the computer rather than in the laboratory we can discover the factors which characterize a "better" cleavage reagent for use in protein-mass fingerprinting.

Setting aside physico-chemical factors, such as percentage of cleavage sites actually cut in a given amount of time, and real-world problems such as sample impurity, proteases differ only in terms of the sets of amino-acids cleaved. For this reason, discussion in this paper will focus on *cutter-sets*, i.e. the sets of amino-acids cleaved by a particular protease, rather than on specific proteases. This change of focus also provides the freedom to experiment on the computer with cutter-sets for which no enzymatic or chemical reagent may currently exist.

The question then becomes, what properties characterize the cutter-sets that are most efficacious for protein-mass fingerprinting? This question can be further refined to three other questions.

- Which cutter sets are more likely to identify an unknown protein or protein fragment?
- Given that certain cutter sets are able to identify an unknown protein, which are able to do so using the fewest fragment molecular masses?
- If more than one cutter is required, what is the most efficacious combination.

2. Materials and Methods

A non-redundant protein database, OWL, was used in this study. (In this context, "non-redundant" means that exact duplicates of proteins already present in the database have been removed.) The June 1995 release contains 128,719 protein sequences ranging in length from 3 to 15,281 amino-acids. The average sequence length is 308.22. Although this database has since been updated, the principles elucidated here will remain the same.

2.1 Assumptions Used When Digesting the Database

Different cutter-sets are then used in turn to digest the proteins in the database, creating for each protein a set of fragments with their fragment masses. That is, for a given cutter-set, each protein in the database is cleaved whenever members of the cutter-set are encountered. This is performed under three assumptions:

2.1.1 Cleavage is always perfect Under certain circumstances certain physical reagents fail to cleave as expected, e.g. when faced with sequences composed of of Phe, Trp or Tyr, chymotrypsin sometimes leaves one or more bonds uncut. In practice, this can be overcome by leaving the digestion to run longer.

2.1.2 Cleavage always occurs on the Carboxyl terminal of an amino-acid While most enzymatic and chemical reagents cleave on the C terminal, a number cleave on the N terminal. However, to make a distinction in these experiments would impart a bias (albeit a small one) in the statistics; fragments generated by N-terminal cutters would be one amino-acid shorter than those generated by C-terminal cutters.

2.1.3 Fragment masses are accurate to the nearest Dalton (i.e. +/- 0.5 Da) While this degree of is often not possible in practice nothing is lost by the added precision; The constraint can be relaxed at a later stage, after an appropriate window size has been determined experimentally.

2.2 Excess Cutters

When a particular cutter-set is applied to the database *excess cutters* pose a special challenge. For example, if the cutter set is E (Glu) and part of the current protein is: ...IRPPLRGQRPQEEEEEEGRHGRHG... then the first E will terminate the fragment ...IRPPLRGQRPQE. The question is what to do with the remaining 5 repetitions of E - the excess cutters. There are three possibilities:

2.2.1 Single Cutters Assuming perfect digestion, each of the excess cutters will form a tiny fragment containing just that amino-acid. Unfortunately, because excess cutters are relatively common there would be a spike in the count of occurrences for that fragment-mass which would distort the statistics.

2.2.2 No Single Cutters A second strategy is simply to ignore any fragment containing a single amino-acid, or alternatively, knowing the molecular masses of the cutters, to ignore fragments with those masses. This strategy avoids the spike in the counts of occurrences.

2.2.3 Pseudo Fragments The third strategy is to construct *pseudo-fragments* - sequences containing only members of the cutter-set. These sequences terminate just before the first non-cutter. In the example above, the first E terminates the previous fragment. The remaining cutters form a pseudo-fragment of length 5. Pseudo-fragments have the advantage that excess cutters are treated in a manner similar to non-cutters thus providing additional information about proteins which would otherwise be lost.

2.3 Determining a Minimal Covering Set of Fragment Masses for a Protein

A *covering set* for a particular protein is a subset of the fragment masses, generated by a given cutter set, which is sufficient to uniquely identify that sequence from all the others in the database. For example, fragment masses 1,440 and 3,545, though they appear in other sequences, only appear together in sequence 1B46_HUMAN. Moreover, in this case the covering set is also minimal, i.e. there are no smaller sets of fragment masses that uniquely identify this protein. However, of greater interest is the *average minimum covering set*, i.e. for all the sequences for which a minimum covering set can be found, this is the average size of the minimum covering sets.

The algorithm for obtaining a guaranteed minimum covering set for each protein (if one exists) can be shown to be NP-complete. Specifically, in Garey and Johnson's survey, (Garey & Johnson, 1979) the minimum set cover problem is listed as NP-complete, with the proof attributed to Karp. Here what is sought is, for each sequence, the minimum number of fragments such that, when the intersection of their container-sets is taken, a singleton set is returned. That is, each container-set generated by fragments in a particular sequence must at least contain the identifier for that sequence, so if a combination of fragments uniquely identifies the sequence, the intersection of the fragments' container sets will return the singleton set containing just that sequence's identifier. (A *container-set* for a particular fragment mass is the set of sequences containing that mass.)

The following greedy algorithms were therefore used:

For each sequence in the database:

Sort the fragments in order of increasing container-set size. This will reveal those sequences which are uniquely identified by a single fragment. Uncut sequences can also be dealt with immediately; either their single fragment-masses are unique and identify the sequence or they are shared and no solution is possible (at least for that cutter-set).

All the container-sets are intersected. If the singleton set does not result, no solution is possible for that sequence. On the other hand, if a solution is found and the solution involves two fragments, the algorithm can stop because a minimal solution has been found.

The container-sets are intersected pairwise. If a pair of container-sets yields the singleton set, the algorithm for that sequence stops. As mentioned above, any such solutions are minimal.

If no solution is found for pairs of fragments, the fragment-pair with the smallest container-set intersection is then intersected with each of the remaining fragments' container sets. This process is repeated until the singleton set results, retaining at each pass the fragment whose intersection with the current combination produces in the greatest reduction in resultant set size. If a singleton set results from the intersection of 3 fragments, that solution must also be minimal.

Solutions involving 4 or more fragments are not necessarily minimal and can possibly be improved. Two algorithms are used to attempt to improve these solutions:

Each of the fragments' container-sets are removed in turn from the intersection. If the resulting intersection remains unchanged (the singleton set), then that fragment is clearly not essential.

The initial fragment set is randomized and the intersection of all the container sets is taken. The number of fragments required for a covering-set is returned.

The complexity of these algorithms is dominated by the $O(mn^2)$ pairwise comparison of container sets, where n is the length of the list of fragment molecular masses (which in turn is proportional to the length of the input sequences), and m is the length of the container-sets being intersected at each step. Minimal covering sets are typically obtained for more than 99% of the sequences for which a covering set exists.

3. Experiments and Results

3.1 Comparing Excess Cutter Methods

To first settle the choice of excess-cutter methodology, three different cutter-sets: C, L and RK were used to digest the protein database for each of the three possible excess-cutter methodologies. Considering the sizes of the average minimum covering sets together with the numbers of unidentified sequences it was noticeable that for each cutter, the choice of method for dealing with excess cutters did not make a great deal of difference. However, Pseudo-fragments generally performed better than No-Single-Cutters, which in turn was better than Single-Cutters. For this reason, subsequent experiments standardized on Pseudo-fragments.

3.2 Comparing Different Cutter Sets

The database was then digested by cutter sets containing the 20 naturally occurring amino-acids taken singly, together with cutter sets corresponding to the proteases Chymotrypsin (FWY), Trypsin (RK) and Glu-C (ED). The Pseudo-fragments method was used in each experiment. The striking observation was that for each of the cutter sets, the size of the average minimum covering set was around 2.0. Furthermore, each of the cutters left some sequences unidentified, ranging from 10,168 (FWY), 10,598 (G) to 26,104 (W), 23,290 (M).

Regression analysis was carried out across the 23 sets of measurements on the various metrics against the *Count of eligible fragments*, the latter a measure of whether the cutter-set contains *common* (i.e. frequent) or *uncommon* cutters. From this is was clear that cutter sets which aim to reduce the number of unidentified protein sequences will have the effect of increasing the size of the covering sets (specifically the minimum covering sets), and vice versa.

3.3 Impact of Restricting the Range of Eligible Fragments

The techniques currently in use for protein-mass fingerprinting are unable to measure the large range of fragment masses revealed by the computer-based experiments. (The largest fragment, due to the cutter C, had a mass of 639,522Da.) The next experiments therefore examined the impact of restricting the range of eligible fragment masses. Three ranges were chosen:

- 0 to Inf, i.e. open range
- 500 to 5,000 (reflecting the range of masses examined by many protein-mass fingerprinting systems)
- 0 to 5,000

The database was digested by the cutter sets C and RK for each of the three fragment-mass size ranges. The impact of restricting the range of eligible fragment masses was particularly evident on the counts of unidentified protein sequences. For RK, a common cutter, removing the top of the range - from 5,000Da upwards - had relatively little impact, increasing the number of unidentified proteins from 10,278 to 11,677. On the other hand, further restricting the range by removing the bottom of the range had a larger impact, increasing the count from 11,677 to 17,581. In other words, for a common cutter, there would appear to be much useful information in the range of masses below 500Da.

By contrast, an uncommon cutter such as C has most of its useful information in the open-ended range of masses above 5,000Da; when the top of the mass size-range is removed the number of unidentified proteins jumps from 22,909 to 62,616, while also removing the bottom of the range only causes a small further increase, to 66,720.

A subsequent experiment used a cutter set of RK and a sliding mass-range-window of 500Da to see which window contains the most information. The window was advanced in increments of 50Da. That is, increasing eligible-fragment-mass ranges where tried, from 50-549, 100-599, 150-649, and so on to 3,000-3,499. It was found that the number of unidentified proteins falls initially, reaching a minimum in the range 350-849 and then rises steadily. At the same time, the size of the average minimum covering sets rises slightly at the outset reaching a peak of 3.39 for the window 250-749 and then falls steadily, reaching a minimum of 0.22 in the last window sampled (3,000-3,499).

To understand why the smaller fragment masses are so useful, the counts of the various masses were examined for the cutter-set RK. What was noticeable was the sharp peak on the counts in the range 500-1,000 Da. In particular, although many masses around 500Da have very high incidences, many others are remarkably uncommon. Thus, used either singly or in combination with more common masses, they may serve to identify a sequence. Just as importantly, there are few gaps in the range of fragment masses (i.e. masses for which no fragment exists), particularly for masses between 200Da and 5,000Da. For fragment masses above 5,000Da, incidences are typically low with large gaps in the range of fragment masses. Therefore, particularly for the common cutter RK and a fixed window size, the probability of having a unique combination is actually lessened above 5,000Da.

3.4 The Problem with Non-Redundant Databases

Knowing now that each cutter-set leaves a number of sequences unidentified, the question became: will a combination of cutter-sets be able to identify all the sequences in the database? The methodology would be identical to that employed for fragments within sequences: test (in linear time) that a solution is possible

by taking the intersection of all the sets of unidentified sequences; the empty set implies that a solution is possible. The next step would then be to find the smallest such combination.

When the intersection of all 23 cutter sets was taken, 899 sequences remained unidentified. A closer examination of the 899 sequences reveals that:

- The smallest are 3 sequences of length 3; 656 (73%) sequences contain 15 amino-acids or less, and 744 (83%) contain 20 amino-acids or less. (The median value is 11, with the mode at 10.)
- Of the remainder, particularly the 167 sequences with 20 or more amino acids, many are more than 90% similar to others in the group. (Many of these matches are greater than 97%.)
- One unusually long sequence, HUMCR1SF41, 2037 aa, appears in the list of unidentified sequences. However, looking at the complete database one finds a sequence HUMCR1SF411 (2486 aa), which totally encompasses the shorter sequence in two substrings and effectively shadows it.

4. Discussion and Conclusions

Despite not being able to name the cutter-sets which can identify every sequence in the OWL database, the investigation has highlighted the characteristics of such a set of cutter-sets.

4.0.1 Minimum of two fragment masses On average, a minimum of two fragments are required for both common and rare cutters to successfully identify sequences, although the latter left more sequences unidentified. Generally, it has become clear that, of the two criteria for a "better" cutter, reducing the total number of unidentified sequences is more important than reducing the number of fragments that must be tested. The average number of peptide fragments generated per sequence by each of the 23 cutters varied between 6.4 for a C-cutter to 30.9 for Glu-C. The combination of peptides chosen for database searches dramatically affects the search outcome, so one possibility is to undertake some form of *combinatorial sieving*. That is, digesting with a particular cutter-set will produce an average of around 20 peptide fragments. Each of these could be tried singly against a database engine such as MOWSE in the hope of identifying candidate proteins. Should this be unsuccessful, the next step could be to query with all possible pairs of fragments. If this also fails an alternate cutter-set should be tried.

4.0.2 More than one cutter-set No cutters were able to uniquely identify all the sequences in the database. Another unexpected result was the very large size of some sequences left intact by even common cutters, e.g. 60,906Da for RK (Trypsin), and ranging out to 258,314Da for C. This finding further strengthens the need for the use of more than one cutter in experimental studies. Different cutter-sets must be tried in succession, noting that combining cutters in the one experiment (if possible chemically) may only have the effect of creating a multiple cutter which is approximately the sum of the input cutter sets. This number of cutters may be quite high if one considers that even when all 23 cutter sets were used 899 of the 128,719 sequences remained unidentified. As database size increases, this figure may represent a significant number of proteins. However, it is conjectured that only a small number of cutters may be required to place each protein sequence greater than some minimal size (e.g. 10 amino-acid residues) within a group of related sequences.

4.0.3 Common versus Rare Cutters It is clear from the range-restriction experiments that common cutters provide much useful information within the lower range of fragment masses. (These values are often ignored in practice.) On the other hand, uncommon cutters such as C (2-Nitro-5-thiocyanobenzoate) or M (Cyanogen bromide) have their most useful information in mass ranges not normally available to protein-mass fingerprinting systems and therefore may be less effective at identifying unknown proteins. In other words, reagents that are common cutters appear to be more useful for PMF than those that are uncommon cutters.

4.0.4 Excess cutter strategies in practice Another conclusion that can be drawn from the experiments with different excess cutter strategies is that, even if pseudo-fragments are not realizable in a laboratory setting, ignoring single cutters is more effective than taking them into account.

4.0.5 More strongly non-redundant databases A non-redundant database is a requirement for these experiments because if two copies of the same protein sequence occur in the database, every fragment of one will also be a fragment of the other and neither sequence will be identified. It is now clear that even this database is not sufficiently non-redundant. That is, two sequences can differ by only a small number of amino-acids and yet both will be in the database. A similar problem arises when one sequence is also a part of a larger sequence in the database. For this reason, large databases of protein sequences need to be constructed, similar to those created by Hobohm et. al (Hobohm et al., 1992), but much larger, where no sequence has no more than N% sequence similarity with any other.

5. References

Garey, Michael R. and Johnson, David S. (1979), *Computers and Intractability: A Guide to the Theory of NP-Completeness.* W. H. Freeman.

Hobohm, Uwe, Houthaeve, Tony and Sander, Chris (1994), Amino Acid Analysis and Protein Database Compositional Search as a Rapid and Inexpensive Method to Identify Proteins. *Analytical Biochemistry* **222**, pp. 202-209.

Hobohm, Uwe, Scharf, Michael, Schneider, Reinhard and Sander, Chris (1992), Selection of Representative Protein Data Sets. *Protein Science* **1**, pp. 409-417.

Pappin, D. J. C., Hojrup, P. and Bleasby, A. J. (1993), Rapid Identification of Proteins by Peptide-Mass Fingerprinting. *Current Biology* **3**, pp. 327-332.