Peptide sequence reconstruction from de novo sequences and their homologues

Denis Yuen, Bin Ma, Iain Rogers

Bioinformatics Solutions Inc.

Introduction

Because protein sequence databases will never be complete, contain gene prediction errors, and can't account for mutations between individuals, it is often necessary to derive a peptide sequence from MS/MS data where no exact match can be found in the database. De novo sequencing provides a useful technique for sequencing peptides without a database, but completely correct sequences are difficult to find. However, when coupled with a sequence tag homology search like SPIDER[1], similar peptides can be returned from a protein sequence database.

Here we present a technique for constructing the real peptide sequences from de novo sequences derived by PEAKS Studio[2] and homologous entries from a database.

Approach

A problem inherent to de novo sequencing is resolving the correct amino acid assignment where more than one amino acid, or combination of amino acids can account for the mass difference between two peaks. Leucine=Isoleucine and K=Q are two commonly cited examples. Similarly, N=GG, AG=Q, and of course TL = LT. These errors, however, are non-critical if they can be easily accounted for in downstream analysis.

The SPIDER search tool, for peptide sequence homology, can account for non-critical de novo sequencing errors in its analysis, enabling researchers to identify homologues for interesting peptides that are not present in any protein sequence database. When comparing the original de novo sequence to the homologue, the problem remains: how to quickly identify which differences constitute legitimate mutations, and which are more likely to be allowable *de novo* sequencing errors. The true sequence must fit somewhere between the de novo sequence and its homologue.

'Sequencing error' can be computed as the minimum number of substitutions required to correct a de novo sequence to a proposed 'better' sequence. Similarly, 'edit distance' is based on the minimum number of insertions, deletions and mutations [weighted by a Blosum 62 matrix] that must be made to edit a homologous sequence to match the proposed 'better' one.

So then, an algorithm is constructed to find the 'better' sequence, that fits between a given de novo sequence and a given homologue, to minimize the edit distance and the sequencing error. But finding the true sequence relies on having a good de novo sequence and a good homologous sequence.

Further then, the algorithm must find the homologue in the database that results in the 'better' sequence having the absolute minimum of edit distance and sequencing error.

Algorithm Theory - Reconstruction

Let

$d_s(X,Y)$ = sequencing error between X and Y

$d_h(Y,Z)$ = homology mutations between Y and Z

Sequencing: Given de novo sequence X, homologue Z, find Y such that $d_s(X,Y)+d_h(Y,Z)$ is minimized.

Let $d(X,Z)=\min_Y d_s(X,Y)+d_h(Y,Z)$

Searching: search a database for Z such that d(X,Z) is minimized.

The core problem is to compute d(X,Z).

Sequencing error cost d_S(X,Y)

Easily align X and Y together (according to mass).

```
(Seq)   X:   LSCFAV
(Real)  Y:   EACFAV
```

For each erroneous mass block (Xi,Yi), define the cost on the block

$$d_s(X_i,Y_i)=f(m(X_i))$$

f(m) depends on how often a same mass replacement error of mass m is observed. The more frequent, the smaller.

Define $d_s(X,Y)=\sum_i f(m(X_i))$

How to Compute d_h(X,Z)

A multiple alignment can be built from alignments (X,Y) and (Y,Z).

```
(Seq)   X:   LSCF-AV
(Real)  Y:   EACF-AV
(Match) Z:   DACFKAV
```

Lemma: $d(X,Z)=\sum_i d(X_i,Z_i)$

So let $D(i,j)=d(X[1..i],Z[1..j])$ and use dynamic programming to solve the rest.

Results

A sample of Bovine Serum Albumin was analyzed on a Waters Q-TOF mass spectrometer. The resulting data was searched, using PEAKS Protein ID, against the Swiss-Prot database to identify 28 peptides that matched exactly to ALBU_BOVIN. De novo sequences, as derived from PEAKS auto de novo, for those 28 spectra were extracted.

The de novo sequences were searched, using SPIDER, against the Human database, where we can expect to find several homologous Albumin peptide sequences, but few exact matches. For each spectrum, the proposed 'better' sequence (as reconstructed from the human homologue and the de novo sequence) was compared back to the true protein, ALBU_BOVIN, to evaluate correctness.

Of the 28 spectra, PEAKS auto de novo gave 13 completely correct and 15 partially correct sequences. SPIDER searching returned reasonable homologous peptides from Human Albumin for all spectra. The algorithm's proposed 'better' sequence was correct (i.e. matched exactly to ALBU_BOVIN) in 24 cases.

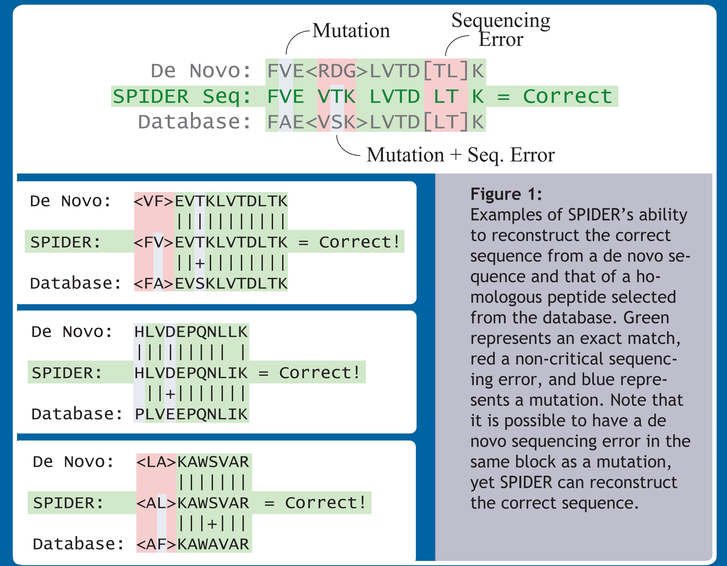

Figure 1: Examples of SPIDER's ability to reconstruct the correct sequence from a de novo sequence and that of a homologous peptide selected from the database. Green represents an exact match, red a non-critical sequencing error, and blue represents a mutation. Note that it is possible to have a de novo sequencing error in the same block as a mutation, yet SPIDER can reconstruct the correct sequence.

Conclusions

SPIDER has a demonstrated ability to identify important mutations in peptide sequences, distinguishing them from de novo sequencing errors. This enables researchers to discover new peptides not present in any database and should prove useful in any protein characterization research or proteomics application.

References

1. Y. Han, B. Ma, and K. Zhang: SPIDER: Software for Protein Identification from Sequence Tags Containing De Novo Sequencing Error. Journal of Bioinformatics and Computational Biology 3(3):697-716. 2005.
2. Ma, B., Zhang, K., Hendrie, C., Liang, C., Li, M., Doherty-Kirby, A., Lajoie, G., PEAKS: Powerful Software for Peptide De Novo Sequencing by MS/MS. Rapid Communications in Mass Spectrometry, 17(20):2337-2342.