Peptide Spectral Library

Peptide Library Design Tools by NeoBioLab

- Overlapping
- Truncation
- T-cell Truncated
- Alanine Scan
- Positional Scan
- Combinatorial 2-Positional Scan
- Combinatorial 3-Positional Scan
- Scrambled

The Peptide Spectral Library refers to an ongoing project that creates a non-redundant and annotated database for all LC-MS/MS peptide spectra. This will provide templates that can be used to identify these proteins during research and allow researchers to more easily collaborate ‘experimental data’, for further scientific studies.

The Methods

The methods used to identify these spectra for identification within the database have also streamlined the ability to approach sequence searching which makes it easier for these items to be readily identified. As the ability to identify these spectra has increased, it has resulted in millions of peptidesbeing documented, increasing the need for a comprehensive database.

As database search and documentation technology has increased, this project has seen feasibility and for the first time, has been acknowledged as a practical endeavor.

Accessibility

To date, a variety of peptide databases have been created. These largely focus on specific types of peptides that may be applied to a variety of fields of research. Some databases have also been created with a field of repetitive entries that work largely to improve the practicality and search ability of these databases so they can be more readily accessed by users.

This helps to ensure that repetitive data is not inadvertently entered and maximizes the search component of these components, to help ensure that users will be able to find the necessary protein and amino acid components of the chemicals they are searching for.

Using an Annotated Library

The system used to create a peptide spectral library requires the system to depend on corresponding sequences within the loaded data the system has catalogued.

- The earliest systems began this by annotating spectra data that was already made available in the Global Proteome Machine Database or GPMDB

- Systems then went a step further and used this library of spectra to create an additional database by averaging any spectra that were annotated within a similar peptide sequence during experimentation. This also includes those that shared a sequence modification or shared a parent ion charge to offer classifications for data within the library.

- The database was structured in a way that experimental peptides with tandem mass spectra could be easily compared with those already listed within the library. This will allow researchers to create a sequence identification method. This method is based on a scoring system that ranks the similarities of experimental spectrum with those that are already established.

Software that implements this system has been constructed to identify these sequence identifiers. Researchers have already built databases for annotated tandem masses– including mus musculus, saccharomyces cerevisiae and Homo sapiens.

These can be readily downloaded by other groups to use in their research and, in most cases, those that access these programs are encouraged to enter their own data into the library, should a discovery that does not match the current established data be discovered.

Methodology for Correlating Mass Spectra Peptides

A new method to help correlate uninterpreted mass spectra peptides, by using low energy collisions with amino acids through the Genpept database, has been developed.

- This methodology will be used to create a protein database that can be searched by the amino acid sequences used within the mass as well as molecular weight.

- This data can be cross correlated with mass to charge rations in fragment ions that can be predicted in amino acid sequences. This information can also be obtained by the additional database of fragment ions in the tandem mass spectrum.

In general, there is a difference of .1 or greater for the normalized cross-correlation functions for first and second ranked search results. This database can manage searches for species specific protein databases as well.

This is utilized by matching the mass spectra that is obtained enzymatically when the peptide is digested with the listing in the database. Users will also be able to match proteins with sequences found in organisms such as S. cerevisiae and E. coli that are commonly used to interpret tandem mass spectra to provide a manuscript that is convenient for the database outline.

COPaKB Database

In addition to a standardized peptide library, the Cardiac Orangellar Protein Atlas Knowledgebase (COPaKB) has worked to create a comprehensive database of cardiovascular instigators.

- This project was initially founded by the NHLBI Porteomics Centers and European Bioinformatics Institute, Scripps Research Institute, Royal Institute of Technology, University of California, Los Angeles, Zhejiang University and Beijing Genomics

Institute.

- The goal of this project is to create a comprehensive listing of cardiac proteome dynamics that can encourage collaborative efforts between individual research entities worldwide.

- These goals are addressed by creating a guide to results of previous studies, following up on results from ongoing analyses that can be combined with previous research and enabling a Wiki component that allows researchers from a variety of backgrounds access to these materials as well as the opportunity to participate in the growth of the project.

Participants in the COPaKB believe that providing information regarding these mass spectral features, genetic signatures, clinical attributes and protein expression imaging will advance the growth of cardiovascular biology.

The more that is known about peptides that affect the cardiovascular system, the easier it will be to apply these items to patients properly to address a wide variety of medical concerns. This data-driven analysis versus research into hypothesis theories is believed to fuel additional discoveries and information that can be shared instantaneously via the web to encourage collaborative research.

Peptide Atlas

Peptide Atlas was founded to create a full annotation of eukaryotic genomes as validated and expressed through a protein structure.

- Founders outlined a set of goals for the first publication of Peptide Atlas which outlines the principles of obtaining information regarding peptide mapping and sequences, as well as the proper way to store this information in a database.

- All protein identifications are annotated and clustered with peptide locations on the chromosomal coordinates calculated via the standard CDS coordinates.

- Protein mixtures for each list has been prepared, labeled purified and, in some cases,

digested using trypsin. These samples will be run through a mass spectrometer and this data will be compared to other theoretical or actual spectra to identify any potential peptides.

- These identified peptides will be formed into true and false positive distributions, scored and filtered so they will only retain the highest scoring identifications.

- These sequences will be compared to any peptides already listed in sequence databases to ensure a unique name can be applied to the given protein and where the protein is designed to start.

The Peptide Atlas works to create a framework and method that can be applied to similar programs and proteomics technology so a consistent framework can be developed to create and identify peptides.

This online database will administer experimental data so it is available in the public domain for those that would like to continue this work. The schema will accommodate different builds for organisms so that the sequence can be adjusted based on the starting material used to develop the protein. Anyone who has data on this subject is encouraged to contribute to the work of the database.

Methods to Compare Collision-Induced Peptides

A method to better compare collision induced dissociation (CID) to create a spectra of peptides has been developed to better understand how to categorize these chemicals.

- Cross-correlation analysis of this spectrum can be used to normalize the cross correlation score, comparing the data to the autocorrelation CID sprectra. This can be used to compare mass information as well as fragmentation patterns of developed chemicals.

- These comparisons are also performed by the instrument type used in the creation of the peptide as well as alternative instrument types and tandem mass spectra that were obtained during these operations. This allows for accurate and reliable comparisons that can be demonstrated with repetitive analysis.

- Scores from the cross comparison of these peptides can be used to identify the chemicals should the same peptide be found in lower concentrations than other comparisons that are performed with a similar spectra.

- This method is also insensitive to variations that can occur with daily calibrations of the instruments used.

If minor variations of the peptide can be found in abundance in a given sample, the comparison that was demonstrated using this method can be used to create a library search for these peptides.

The subtractive analysis of the analysis has been found to be particularly effective in managing the tandem mass spectra experiments that include LC/MS/MS chemicals.

Building Consensus Spectral Libraries

One inefficiency that has been noticed with proteomics experiments is the rediscovery of similar peptides that limit the sequence of data searching, leading to an increase in the likelihood of errors.

- In order to develop a more efficient method to identify peptides, is to work through the MS/MS spectra to catalogue and condense these chemicals into a spectral library with a searchable component so new identifications can easily be matched to existing peptides to offer positive alternatives for items that do not appear to have a strong medical application.

- An open source library is already functionally complete and has been applied to an extensible MS/MS search tool. This program is known as SpectraST and offers high quality library search technology that offers a high confidence identification of millions of spectra. These spectra can be identified from a variety of searches and searched using four search engines within the website. Currently the spectra for Saccharomyces cerevisiae contains around 30,000 identified spectra.

- This system has been found to perform at a higher level than the established peptide search engine SEQUEST. It is both faster and able to eliminate a larger number of bad

hits to ensure a smooth user experience.

SpectraST has fully integrated Tarns Proteomic Pipeline software that allows the user to take advantage of essential functionalities such as assigning probability to peptide and proteins, data visualization and quantification of peptides. This program has been specifically designed to target proteomic applications which help to increase the overall functionality of the product.

Potential Shortcomings

The peptide fragmentation process is quite complex which can create difficulties when working to identify all of the items created within this spectrum.

- Using a mass spectra for protein identification can help to calculate all of the potential putative peptide candidates in a setting. However the search engines will be required to apply a variety of heuristics in order to predict fragmentation patters of peptide candidates within their database. This makes it difficult to determine the best possible template that can be used to identify experimental spectra within the protein and peptide identification database. Mascot and SEQUEST have attempted to answer this query in the past but more work is needed to move these efforts forward.

- In some cases a peptide may fragment and fail to produce patters that can be reproduced to create mass spectra or distinct fragments. Because of this, the system may face a bottleneck of specificity that can limit the ability to list potential data from experimental systems.

Because so many of the existing databases are quite large, it can make the process of searching through these databases for information somewhat slow. It may require a high-performance computer to search the database in the first place which can limit access to this information, particularly for up and coming laboratories.

Some have also found that the setup of the Sequence Database Searching mechanism may cause it to disconnect discoveries in a research setting from groups that are already established in the program, providing inconsistent search results for those attempting to use the program. As search capabilities are researched further, these issues are becoming less frequent with newer applications of this technology.

Much of the concern regarding the establishment of a concise peptide spectral library falls upon the idea that the system would be able to account for the millions of entries that it would need to hold over time without inadvertently creating repetitive or unnecessary data.

The Peptide Spectral Library works to integrate programs that will streamline the search capabilities of the program so that contributors can ensure they are not creating a repetitive entry.

Sources:

‘‘COPaKB.” *Cardiac Organellar Protein Atlas Knowledgebase (COPaKB)*. N.p., n.d. Web. 26 July 2013. <http://www.heartproteome.org/copa/default.aspx>.

Craig, R., J.C. Cortens, D. Fenco, and R.C. Beavis. “Using Annotated Peptide Mass Spectrum Libraries for Protein Identification.” *Journal of Proteome Research*. ACS Publication, n.d. Web. 26 July 2013. <http://pubs.acs.org/doi/abs/10.1021/pr0602085>.

Eng, Jummy K., Ashley L. McCormack, and John R. Yates, III. “An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database.” *Science Dirrect*. SciVerse, n.d. Web. 26 July 2013. <http://www.sciencedirect.com/science/article/pii/1044030594800162>.

Lam, Henry, Eric W. Deutsch, James S. Eddes, Jimmy K. Eng, Nichole King, Stephen E. Stein, and Ruedi Aebersold. “Development and Validation of a Spectral Library Searching Method for Peptide Identification from MS/MS.” *Proteomics*. Online Library, 13 Feb. 2007. Web. 26 July 2013. <onlinelibrary.wiley.com/doi/10.1002/pmic.200600625/full>.

‘‘Peptide Atlas.” *Peptide Atlas*. N.p., n.d. Web. 26 July 2013. <http://www.peptideatlas.org/overview.php>.

Yates, John R., III, Scott F. Morgan, Christine L. Gatlin, Patrick R. Griffen, and Jimmy K. Eng. “Method To Compare Collision-Induced Dissociation Spectra of Peptides:? Potential for Library Searching and Subtractive Analysis.” *Analytical Chemistry*. ACS Publication, n.d. Web. 26 July 2013. <http://pubs.acs.org/doi/abs/10.1021/ac980122y>.