Lecture 2

1. Explain, why protein sequences are cleaved with enzymes?
What enzymes do you know, which are the most used?
How do they work, and when do they not work?

Объясните, почему белковые последовательности расщепляются ферментами? Какие ферменты вы знаете, какие наиболее используемые? Как они работают и когда они не работают?

Protein cleavage

- It is needed
- to reduce the complexity and
- to decrease the size of the molecules
- because instruments cannot deal with large molecules.

trypsin, chymotrypsin, and pepsin

Enzymes

Enzymatic cleavage works by breaking the peptide bonds between amino acid residues in a protein sequence.

- Enzymes cut proteins at certain positions.
- Different enzymes cleave at different sites.

Inaccessible sites

- Enzymes may not cleave at certain sites if it is inaccessible because of the spatial structure of the protein.

+ pH, temperature, and concentration of the enzyme and substrate.

Расщепление белков

- Это необходимо
- для уменьшения сложности и
- уменьшить размер молекул, потому что приборы не могут работать с большими молекулами.

Ферменты

Ферментативное расщепление происходит путем разрыва пептидных связей между аминокислотными остатками в последовательности белка.

- Ферменты разрезают белки в определенных местах.
- Разные ферменты расщепляют в разных местах.

Недоступные участки

- Ферменты могут не расщеплять определенные участки, если они недоступны из-за пространственной структуры белка.

2. What is in silico cleavage, how is it defined for example?
Что такое расщепление in silico, как оно определяется, например?

• Knowing the cleavage sites protein sequences can be digested in silico.

- Зная места расщепления, белковые последовательности могут быть переварены
silico.

In silico cleavage — computational prediction of the cleavage sites in a protein sequence by specific enzymes. It is a bioinformatics approach that can be used to simulate enzymatic cleavage and predict the resulting peptide fragments.

For example, in silico trypsin digestion involves identifying all the lysine and arginine residues in a protein sequence and predicting the cleavage sites based on the specificity of trypsin.

Расщепление in silico относится к вычислительному прогнозированию сайтов расщепления в белковой последовательности специфическими ферментами. Это биоинформационный подход, который может быть использован для моделирования ферментативного расщепления и прогнозирования получаемых пептидных фрагментов.

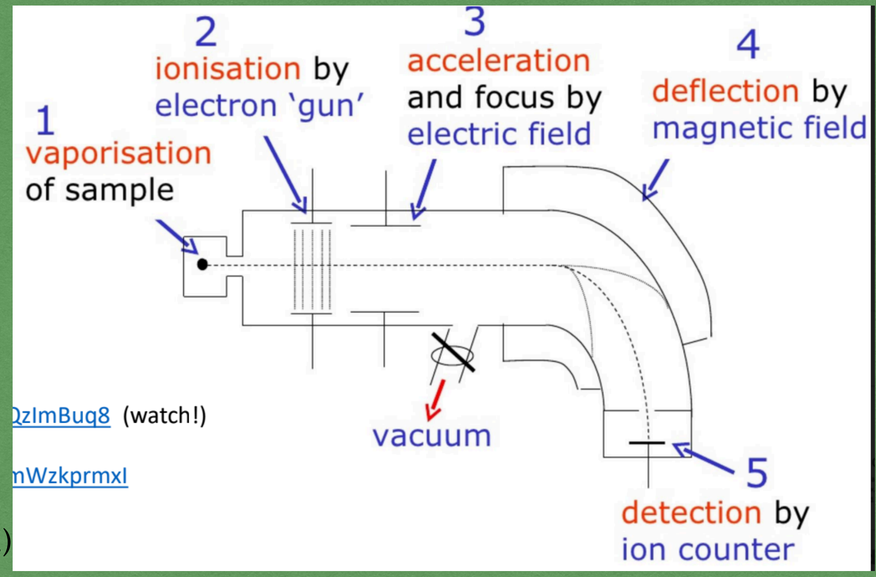
Например, трипсиновое переваривание in silico включает в себя определение всех остатков лизина и аргинина в белковой последовательности и предсказание мест расщепления на основе специфичности трипсина.

3. Describe the schematic parts of a mass spectrometer.

Опишите схематические части масс-спектрометра.

Mass Spectrometry
(Mass. Spec. or MS) · Mass Spectrometer is the instrument which takes an aliquot of molecules as input and produces a spectrum (a mass distribution) of molecular masses or the mass of the fragments.

Масс-спектрометр - это прибор.
который принимает аликвоту молекул в качестве
и производит спектр (распределение масс
распределение) молекулярных масс или
масс фрагментов.



QzImBuq8 (watch!)

mWzkprmxI

The main components of a mass spectrometer are:

- Inlet system (LC, GC, Direct probe etc...)
- Ion source (EI, CI, ESI, APCI, MALDI, etc...)
- Mass analyzer (Quadrupole, TOF, Ion Trap, Magnetic Sector)
- Detector (Electron Multiplier, Micro Channel Plates MCPs)

Основными компонентами масс-спектрометра являются:
Входная система (LC, GC, прямой зонд и т.д...)
Источник ионов (EI, CI, ESI, APCI, MALDI и т.д...)
Масс-анализатор (квадрупольный, TOF, ионная ловушка, магнитный сектор)
Детектор (электронный умножитель, микроканальные пластины MCP)

The instrument consists of three major components:

1. Ion Source: For producing gaseous ions from the substance being studied.
2. Analyzer: For resolving the ions into their characteristics mass components according to their mass-to-charge ratio.
3. Detector System: For detecting the ions and recording the relative abundance of each of the resolved ionic species.

The pipeline of the process:
1. Vaporization
2. Ionization
3. Acceleration
4. Deflection
5. Detection
· The process is carried out in vacuum (collision with other molecules could harm the ions)

Конвейер процесса:
1. Испарение
2. Ионизация
3. Ускорение
4. Отклонение
5. Обнаружение
- Процесс осуществляется в вакууме (столкновение с другими молекулами может повредить ионы)

Прибор состоит из трех основных компонентов:
Источник ионов: Для получения газообразных ионов из исследуемого вещества.
Анализатор: Для разделения ионов на их характерные массовые компоненты в соответствии с их отношением массы к заряду.
Детекторная система: Для обнаружения ионов и регистрации относительной массы каждого из разрешенных ионных видов.

3. Describe the schematic parts of a mass spectrometer.

Опишите схематические части масс-спектрометра.

Vaporization – part I

• Often heats up the matter to transform it to gas phase.
• Less useful with fragile or complex molecules because they might break
• Whether it is required or not, it depends on the ionization procedure.

Испарение - часть I

- Часто нагревает вещество, чтобы перевести его в газовую фазу.
- Менее полезно при работе с хрупкими или сложными молекулами, поскольку они могут разрушиться.
- Требуется или нет, зависит от процедуры ионизации.

Acceleration – part III

• Now, sample molecules are ionized, called parent ions.
• Acceleration speeds up the parent ions.
• It resembles a particle accelerometer

Ускорение - часть III

- Теперь молекулы образца ионизированы и называются родительскими ионами.
- Ускорение ускоряет родительские ионы.
- Это напоминает акселерометр частиц

Deflection – part IV

• In this step the propagation path of charged particles are deflected by a magnet.
• Heavier molecules are deflected less.
• Lighter molecules are deflected more.
• This process will separate (sort) charged particles according to their mass-per-charge ratio.

Отклонение - часть IV

- На этом этапе пути распространения заряженных частиц отклоняются с помощью магнитом.
- Более тяжелые молекулы отклоняются меньше.
- Более легкие молекулы отклоняются сильнее.
- Этот процесс отделяет (сортирует) заряженные частицы в соответствии с их соотношению массы и заряда.

4. What is the ionization step for, how does it work, and what are the types of that? What are the benefits and disadvantages of different ionization procedures?

Для чего нужна ступень ионизации, как она работает и какие бывают ее виды? Каковы преимущества и недостатки различных процедур ионизации?

Ionization – part II

• In this step the molecules become positively charged (kation).
• Charged molecule is called molecular ion or parent ion or precursor ion.
• Molecular ions can become unstable: when it breaks, its parts are called fragment ions.

• Mainly, there are two ways to charge molecules:
• Electron(s) removed from molecules.
• Molecular ion lose a little mass (hardly or not detectable mass change), the mass/charge ratio does not change significantly

• An extra proton is added to the molecules.
• Molecular ion gains a weight of a proton, mass/charge ratio changes considerably.
• The mass of a proton is 1.00727646688 Da

Ionization techniques

• Electron Ionization
• Chemical Ionization
• Atmosphere Pressure Ionization
• Electrospay Ionization (ESI)
• Thermospray
• Matrix Assisted Laser desorption ionization (MALDI)
• Collision-Induced-Dissociation (CID)

Ионизация - часть II

- На этом этапе молекулы становятся положительно заряженными (катион).
- Заряженная молекула называется молекулярным ионом, или родительским ионом, или ионом-предшественником.
- Молекулярный ион может стать нестабильным: когда он разрушается, его части называются
фрагментные ионы.

- В основном, существует два способа зарядить молекулы:
- Электрон(ы) удаляется из молекул.
- Молекулярный ион теряет немного массы (едва или не обнаруживаемое изменение массы), отношение масса/заряд
существенно не меняется

- К молекулам добавляется дополнительный протон.
- Молекулярный ион приобретает массу протона, соотношение масса/заряд значительно изменяется.
- Масса протона составляет 1,00727646688 Да

Методы ионизации

- Электронная ионизация
- Химическая ионизация
- Ионизация при атмосферном давлении
- Электрораспылительная ионизация (ESI)
- Термоспрей
- Лазерная десорбционная ионизация с матричной поддержкой (MALDI)
- Коллизионно-индуцированная диссоциация (CID)

4. What is the ionization step for, how does it work, and what are the types of that? What are the benefits and disadvantages of different ionization procedures?

Для чего нужна ступень ионизации, как она работает и какие бывают ее виды? Каковы преимущества и недостатки различных процедур ионизации?

Chemical Ionization

- Chemicals such as ammonium NH_3 xor methane CH_4 are very potential proton donors. They will pass a proton to the sample molecules.
- Molecular ion charged with extra proton is more stable than when it is charged by removing an electron.
- Disadvantage is that heating may harm large molecules.

Химическая ионизация

- Химические вещества, такие как аммоний NH_3 х или метан CH_4 являются очень потенциальными донорами протонов. Они будут передавать протон к молекулам образца.
- Молекулярный ион, заряженный дополнительным протоном, более стабилен, чем когда он заряжен путем удаления электрона.
- Недостатком является то, что нагревание может повредить крупные молекулы.

Soft ionization

- It does not utilize electrons or chemical molecules for ionization
- It does not need heating, can be used with fragile molecules
-

Мягкая ионизация

- Для ионизации не используются электроны или химические молекулы
- Не требует нагрева, может использоваться с хрупкими молекулами

MALDI - Matrix Assisted Laser Desorption Ionization

- Soft ionization
- Better for larger molecules such as peptides
- Requires vacuum

MALDI - лазерная десорбция с матричной поддержкой Ионизация

- Мягкая ионизация
- Лучше для больших молекул, таких как пептиды
- Требуется вакуум

Electrospray Ionization (ESI)

- Used for large molecules such as peptide or DNA parts. Does not require vacuum. Source is paced in N_2 gas

Ионизация электрораспылением (ESI)

- Используется для больших молекул, таких как пептиды или части ДНК. Не требует вакуум. Источник помещается в N_2 газ

5. Explain the resolution of a detector. How is it measured. What is one Dalton?

Объясните разрешающую способность детектора. Как она измеряется. Что такое один дальтон?

Detector– part V

• Detects charged particles. Scale:
Mass/Charge (m/z)
• Resolution is high enough to
distinguish between isotopes nowadays.
• R depends on the mass.

Detector

• Mass is measured by Daltons.
• Carbon is used as a reference, and 1 Da is defined as one-twelfth of carbon (C/12).
• The mass of carbon is 12.000000 Da.
• The mass of a hydrogen is: 1.007825035
• Low-resolution mass spectrometer's accuracy is within 1 Dalton.
• New, high-resolution mass spectrometer's accuracy is around 0.02 Da
• That is, it can make a difference between the 1/50 of the mass of a proton.

The resolution of a detector in mass spectrometry refers to its ability to distinguish between ions with small differences in their mass-to-charge ratio (m/z). It is measured by the full width at half maximum (FWHM) of the peak in the mass spectrum.

One Dalton (Da) is a unit of mass equal to one twelfth of the mass of a carbon-12 atom. It is commonly used in mass spectrometry to express the mass of ions or molecules. For example, the molecular weight of water (H2O) is approximately 18 Da.

Детектор - часть V

- Обнаруживает заряженные частицы. Шкала:
Масса/заряд (m/z)
- Разрешение высокое достаточно, чтобы различать между изотопами в настоящее время.
- R зависит от массы.

Детектор

- Масса измеряется в дальтонах.
- Углерод используется в качестве эталона, и 1 Да определяется как одна двенадцатая часть углерода
(C/12).
- Масса углерода составляет 12,000000 Да.
- Масса водорода составляет: 1,007825035.
- Точность масс-спектрометра низкого разрешения находится в пределах 1 дальтона.
- Точность нового масс-спектрометра высокого разрешения составляет около 0,02 Да.
- То есть, он может сделать разницу между 1/50 массы протона.

Разрешение детектора в масс-спектрометрии относится к его способности различать ионы с небольшими различиями в их отношении массы к заряду (m/z). Оно измеряется полной шириной на половине максимума (FWHM) пика в масс-спектре.

Один дальтон (Da) - единица массы, равная одной двенадцатой массы атома углерода-12. Она обычно используется в масс-спектрометрии для выражения массы ионов или молекул. Например, молекулярная масса воды (H2O) составляет приблизительно 18 Da.

6. Explain the two pass MS/MS mass spectrometer.

MS/MS

• MS/MS means using two mass analyzers (combined in one instrument).
• Two phase method:
• First, an MS analysis is carried out to measure precursor ion mass-to-charge,
• Then in the second MS phase, the spectrometer selects ions with a specific range of the mass/charger ratio of the precursor ion and generates fragment ions of the precursor ion to give a distribution of the mass of the fragment particles. Parent ions are broken in a gas chamber.

The two pass MS/MS mass spectrometer is a type of tandem mass spectrometer that consists of two stages of mass analysis. In the first stage, the ions of interest are selected based on their mass-to-charge ratio (m/z) using a mass filter such as a quadrupole or an ion trap. These selected ions are then fragmented by collision with a neutral gas such as helium or nitrogen in a collision cell.

In the second stage, the resulting fragments are analyzed based on their m/z values using a second mass filter. This second mass filter can be the same as the first one or a different type such as a time-of-flight (TOF) analyzer. The resulting mass spectrum provides information about the structure and composition of the original ions.

The two pass MS/MS mass spectrometer offers several advantages over single-stage mass spectrometry. It allows for more selective and sensitive detection of target compounds by reducing interference from other ions in the sample. It also provides more detailed structural information about the target compounds by analyzing their fragmentation patterns.

МС/МС

- MS/MS означает использование двух масс-анализаторов (объединенных в одном приборе).
- Двухфазный метод:
- Сначала проводится МС-анализ для измерения массы-заряда ионов-предшественников,
- Затем на второй фазе МС спектрометр отбирает ионы с определенным диапазоном отношения масса/заряд иона-предшественника и генерирует фрагментные ионы иона-предшественника для получения распределения масс фрагментарных частиц. частиц. Родительские ионы разбиваются в газовой камере.

Двухпроходной масс-спектрометр MS/MS - это тип тандемного масс-спектрометра, который состоит из двух стадий масс-анализа. На первом этапе интересующие ионы отбираются на основе их отношения массы к заряду (m/z) с помощью масс-фильтра, такого как квадруполь или ионная ловушка. Затем эти выбранные ионы фрагментируются путем столкновения с нейтральным газом, таким как гелий или азот, в коллизионной ячейке.

На втором этапе полученные фрагменты анализируются на основе их значений m/z с помощью второго масс-фильтра. Этот второй масс-фильтр может быть таким же, как и первый, или другого типа, например, времяпролетный (TOF) анализатор. Полученный масс-спектр дает информацию о структуре и составе исходных ионов.

Двухпроходный масс-спектрометр MS/MS обладает рядом преимуществ перед одноступенчатой масс-спектрометрией. Он позволяет более избирательно и чувствительно определять целевые соединения за счет уменьшения помех от других ионов в образце. Он также позволяет получить более подробную структурную информацию о целевых соединениях путем анализа их фрагментации.

7. Explain the peptide fragmentation

Peptide fragmentation is a process in mass spectrometry where peptides are broken down into smaller fragments by applying energy to the peptide ions. The energy causes the peptide bonds to break, resulting in the formation of fragment ions. These fragment ions can be analyzed to determine the amino acid sequence of the original peptide, as well as any modifications or structural features present. The fragmentation process can occur through different methods, such as collision-induced dissociation (CID) or higher-energy collisional dissociation (HCD), which can result in different types of fragment ions. Peptide fragmentation is an important tool in proteomics research for identifying and characterizing proteins based on their peptide fragments.

Фрагментация пептидов - это процесс в масс-спектрометрии, при котором пептиды разбиваются на более мелкие фрагменты путем приложения энергии к ионам пептидов. Под действием энергии происходит разрыв пептидных связей, что приводит к образованию фрагментных ионов. Эти фрагментные ионы могут быть проанализированы для определения аминокислотной последовательности исходного пептида, а также любых модификаций или структурных особенностей. Процесс фрагментации может происходить с помощью различных методов, таких как диссоциация, вызванная столкновениями (CID) или коллизионная диссоциация с более высокой энергией (HCD), что может привести к различным типам фрагментных ионов. Фрагментация пептидов является важным инструментом в исследованиях протеомики для идентификации и характеристики белков на основе их пептидных фрагментов.

Lecture 3

8. Compare the DeNovo sequencing against database searching. What are their benefits and disadvantages?

Сравните секвенирование DeNovo с поиском в базе данных. В чем их преимущества и недостатки?

Comparison of DeNovo vs Database Searching

• **De Novo:**
• Pros: works with newly sequenced organisms
• Cons: Difficult to validate the results.

• **Database Search (DS):**
• Pros:
• Simple and straightforward
• Has a limited search space.
• Completeness
• Statistical analysis can be carried out.
• Cons:
• Has a limited search space. Limited to the database.
• Longer searching time

Сравнение DeNovo с поиском в базах данных

- De Novo:
- Плюсы: работает с недавно секвенированными организмами
- Минусы: трудно подтвердить результаты.

- Поиск в базе данных (DS):
- Плюсы:
- Простой и понятный.
- Имеет ограниченное пространство поиска.
- Полнота
- Можно проводить статистический анализ.
- Минусы:
- Имеет ограниченное пространство поиска. Ограничен базой данных.
- происходит слишком медленно

Lecture 3

8. Compare the DeNovo sequencing against database searching. What are their benefits and disadvantages?

Сравните секвенирование DeNovo с поиском в базе данных. В чем их преимущества и недостатки?

DeNovo sequencing involves the reconstruction of the amino acid sequence of a protein based on the fragmentation pattern of its peptides. This approach does not rely on a pre-existing protein database and can identify novel or modified proteins. However, DeNovo sequencing requires high-quality spectra and can be time-consuming and computationally intensive.

On the other hand, database searching involves comparing the experimental mass spectra to a pre-existing protein database to identify matches. This approach is faster and less computationally intensive compared to DeNovo sequencing. However, it relies on the completeness and accuracy of the database, and may not be able to identify novel or modified proteins that are not present in the database.

The benefits and disadvantages of DeNovo sequencing and database searching can be summarized as follows:
DeNovo sequencing:
Benefits:
- Can identify novel or modified proteins
- Does not rely on a pre-existing protein database
- Can provide more accurate results for complex samples
Disadvantages:
- Requires high-quality spectra
- Can be time-consuming and computationally intensive
- May not be able to identify low abundance peptides

Database searching:
Benefits:
- Faster and less computationally intensive compared to DeNovo sequencing
- Can identify proteins with high confidence if a comprehensive and accurate database is available
- Can identify post-translational modifications if they are present in the database
Disadvantages:
- Relies on the completeness and accuracy of the protein database
- May miss novel or modified proteins that are not present in the database
- May produce false positives or false negatives depending on the quality of the spectra and search parameters.

Секвенирование DeNovo предполагает реконструкцию аминокислотной последовательности белка на основе картины фрагментации его пептидов. Этот подход не опирается на существующую базу данных белков и позволяет выявлять новые или модифицированные белки. Однако секвенирование DeNovo требует высококачественных спектров и может быть трудоемким и вычислительно затратным.

С другой стороны, поиск по базе данных предполагает сравнение экспериментальных масс-спектров с уже существующей базой данных белков для выявления совпадений. Этот подход быстрее и менее трудоемок по сравнению с секвенированием DeNovo. Однако он зависит от полноты и точности базы данных и может быть не в состоянии идентифицировать новые или модифицированные белки, которых нет в базе данных.

Преимущества и недостатки секвенирования DeNovo и поиска в базе данных можно суммировать следующим образом:
Секвенирование DeNovo:
Преимущества:
- Может идентифицировать новые или модифицированные белки
- Не полагается на уже существующую базу данных белков
- Может обеспечить более точные результаты для сложных образцов
Недостатки:
- Требуются высококачественные спектры
- Может занимать много времени и требовать больших вычислительных затрат
- Может быть не в состоянии идентифицировать пептиды с низким содержанием.

Поиск в базе данных:
Преимущества:
- Быстрее и менее трудоемкий по сравнению с секвенированием DeNovo
- Может идентифицировать белки с высокой степенью достоверности, если имеется полная и точная база данных
- Может идентифицировать посттрансляционные модификации, если они присутствуют в базе данных
Недостатки:
- Зависит от полноты и точности базы данных белков
- Может пропустить новые или модифицированные белки, которые не присутствуют в базе данных
- Может давать ложноположительные или ложноотрицательные результаты в зависимости от качества спектров и параметров поиска.

9. What is in silico peptide fragmentation, how does it work?

In silico peptide fragmentation is a computational method used to predict the fragmentation patterns of peptides in mass spectrometry experiments. It involves simulating the process of peptide fragmentation by breaking the peptide bonds at specific sites and predicting the resulting fragment ions based on their mass-to-charge ratio (m/z) values.

The process of in silico peptide fragmentation typically involves three steps: (1) selecting a fragmentation method, such as collision-induced dissociation (CID) or higher-energy collisional dissociation (HCD), (2) generating theoretical fragment ions by breaking the peptide bonds at specific sites, and (3) calculating the m/z values of the fragment ions based on their chemical composition.

In silico peptide fragmentation can be useful in proteomics research for identifying and characterizing proteins based on their peptide fragments. By comparing the predicted fragmentation patterns of peptides with experimental data, researchers can confirm the identity of proteins and identify post-translational modifications or other structural features.

Пептидная фрагментация in silico - это вычислительный метод, используемый для предсказания характера фрагментации пептидов в масс-спектрометрических экспериментах. Он включает моделирование процесса фрагментации пептидов путем разрыва пептидных связей в определенных местах и предсказание получаемых фрагментных ионов на основе их отношения массы к заряду (m/z).

Процесс фрагментации пептидов in silico обычно включает три этапа: (1) выбор метода фрагментации, например, столкновительно-индуцированной диссоциации (CID) или высокоэнергетической столкновительной диссоциации (HCD), (2) генерация теоретических фрагментных ионов путем разрыва пептидных связей в определенных местах, и (3) расчет m/z значений фрагментных ионов на основе их химического состава.

Пептидная фрагментация in silico может быть полезна в исследованиях протеомики для идентификации и характеристики белков на основе их пептидных фрагментов. Сравнивая предсказанные схемы фрагментации пептидов с экспериментальными данными, исследователи могут подтвердить идентичность белков и выявить посттрансляционные модификации или другие структурные особенности.

10. What are the components of database-searching?

Key components of the Database-Searching

- Scoring Function
- Protein Database

In this approach an experimental spectrum s is iteratively compared and scored against a large database of reference peptides hjs. The experimental spectrum s is annotated by the best-scoring reference peptide ˆ.
This consists of three key elements: (1) the peptide database DB, (2) a selection of biologically/chemically plausible peptides, called candidate peptides (CP (s) ⊆ DB) with respect to an experimental spectrum s, and (3) the score function $\phi : S \times DB \rightarrow R$, where S denotes the set of spectra obtained from an experiment and R denotes the real numbers. The elements and the size of the CP highly varies for different experimental spectra. The scoring typically provides a similarity-like score (i.e. higher score indicates a better match) based on matching the peaks of the experimental to reference peaks generated from the peptide sequences in silico [3].

В этом подходе экспериментальный спектр s итер-. последовательно сравнивается и оценивается с большой базой данных эталонных пептидов hjs. Экспериментальный спектр s аннотируется эталонным пептидом ˆ, получившим наилучшую оценку. пептидом ˆ.
Эта система состоит из трех ключевых элементов: (1) база данных пептидов DB, (2) выбор биологически/химически правдоподобных пептидов, называемых пептидами-кандидатами (CP (s) ⊆ DB) относительно экспериментального спектра s, и (3) функция оценки $\varphi : S \times DB \rightarrow R$, где S обозначает множество спектров, полученных из эксперимента, а R обозначает вещественные числа. Элементы и элементы и размер CP сильно варьируются для различных экспериментальных спектров. Скоринг обычно дает оценку, подобную оценке сходства (т.е. более высокий балл указывает на лучшее совпадение) на основе сопоставления экспериментальных пиков с эталонными пиками полученными из пептидных последовательностей in silico [3].

11. What is the parent mass tolerance? What is PPM, how does it defined?

Parent mass error tolerance

• In Dalton.
• The error tolerance between the precursor mass and the calculated mass of the theoretical peptide is expressed simply in Dalton.
• Tolerance is given as: ±0.5 Da
• In PPM (Parts per Million):
• MPA: neutral mass of the precursor ion
• Tolerance: δ ppm
• Lower bound: MPA-MPA* δ /1000000
• Upper bound: MPA+MPA* δ /1000000
• Here the amount of the tolerance depends on the actual mass.
• For smaller peptides the tolerance is tighter,
• For larger peptides the tolerance is looser.
• Usually tolerance values for modern instruments are ~10-100 ppm

The parent mass tolerance is the maximum allowed difference between the observed mass of a peptide ion and its theoretical mass.
PPM is a relative measure of mass accuracy that takes into account the size of the parent ion. It is defined as the difference between the observed mass and the theoretical mass, divided by the theoretical mass, multiplied by one million.

Допустимая погрешность родительской массы

- В Дальтонах.
- Допустимая ошибка между массой предшественника и рассчитанной массой теоретического пептида выражается просто в Дальтонах.
- Допуск приводится в виде: ±0,5 Da
- В PPM (частях на миллион):
- MPA: нейтральная масса иона-предшественника.
- Допуск: δ ppm
- Нижняя граница: MPA-MPA* δ /1000000
- Верхняя граница: MPA+MPA* δ /1000000
- Здесь величина допуска зависит от фактической массы.
- Для небольших пептидов допуск более жесткий,
- Для больших пептидов допуск меньше.
- Обычно значения допусков для современных приборов составляют ~10-100 ppm

Допуск родительской массы - это максимально допустимая разница между наблюдаемой массой пептидного иона и его теоретической массой.
PPM - это относительная мера точности массы, которая учитывает размер родительского иона. Он определяется как разница между наблюдаемой массой и теоретической массой, деленная на теоретическую массу и умноженная на миллион.

Lecture 4

12. What kind of experimental spectrum – theoretical spectrum scoring functions do you know. How do they work?

Score functions

Score functions are essentially based on matching experimental peaks to theoretical peaks. Standard score functions are:

Shared Peak Count (SPC)

This is the number of the peaks in the theoretical spectrum that are matched to peaks in the experimental spectrum

Inner product (I)

This is the sum of the intensities of the peaks in the experimental spectrum that match to peaks in the theoretical spectrum. This method takes into account peak intensities. This assumes that noise peaks obtain smaller intensities

XCorr: Cross-correlation function is a widely used scoring function in database searching. It calculates the similarity between the experimental and theoretical spectra by comparing their intensity values at each m/z value.

1. HyperScore: HyperScore is a scoring function used in the SEQUEST algorithm for peptide identification. It calculates the similarity between the experimental and theoretical spectra by comparing their intensity values and taking into account the presence of isotopic peaks.

2. Andromeda Scoring Function: Andromeda is a scoring function used in the MaxQuant software for peptide identification. It uses a probabilistic model to estimate the probability of a match between the experimental and theoretical spectra, taking into account factors such as peak intensity, mass accuracy, and fragmentation pattern.

3. OMSSA: OMSSA is a scoring function used in the Open Mass Spectrometry Search Algorithm for peptide identification. It uses a statistical model to estimate the probability of a match between the experimental and theoretical spectra, taking into account factors such as peak intensity, mass accuracy, and peptide length.

Балльные функции

Функции оценки по существу основаны на сопоставлении экспериментальных пиков с теоретическими пиками. Стандартными функциями оценки являются:

Общее количество пиков (SPC)

Это количество пиков в теоретическом спектре, которые совпадают с пиками в экспериментальном спектре.

Внутреннее произведение (I)

Это сумма интенсивностей пиков в экспериментальном спектре, которые совпадают с пиками в теоретическом спектре. Этот метод учитывает интенсивность пиков. При этом предполагается, что шумовые пики имеют меньшую интенсивность.

XCorr: функция кросс-корреляции - широко используемая функция оценки при поиске в базах данных. Она рассчитывает сходство между экспериментальным и теоретическим спектрами путем сравнения их значений интенсивности при каждом значении m/z.

1. HyperScore: HyperScore - это функция подсчета баллов, используемая в алгоритме SEQUEST для идентификации пептидов. Она рассчитывает сходство между экспериментальным и теоретическим спектрами путем сравнения их значений интенсивности и с учетом наличия изотопных пиков.

2. Скоринговая функция Andromeda: Andromeda - это функция подсчета баллов, используемая в программе MaxQuant для идентификации пептидов. Она использует вероятностную модель для оценки вероятности совпадения экспериментального и теоретического спектров, принимая во внимание такие факторы, как интенсивность пиков, точность определения массы и характер фрагментации.

3. OMSSA: OMSSA - это функция подсчета баллов, используемая в алгоритме Open Mass Spectrometry Search Algorithm для идентификации пептидов. Она использует статистическую модель для оценки вероятности совпадения экспериментального и теоретического спектров, принимая во внимание такие факторы, как интенсивность пика, точность определения массы и длина пептида.

13. What is the discriminative property of scoring functions?

Discriminative power of score functions. It means the ability of the score function to distinguish between the correct and incorrect peptide-spectrum-matches (PSMs).

Дискриминантная способность балльных функций. Она означает способность функции оценки различать правильные и неправильные пептидные спектральными совпадениями (PSM).

The discriminative ability of a scoring functions means that the scores of a correct PSM is much higher, and therefore it is well separated from the score distribution of the incorrect PSMs (i.e. the null distribution); therefore, the correct PSMs can be separated from incorrect ones using a simple threshold.

Unfortunately, the score functions in spectrum identification are hindered by (a) the presence of many unexplained peaks, which stem from the unusual fragmentation of the peptide or contaminating molecules, or (b) the lack of expected fragmentation ions, which fail to be observed in the mass spectrometer [15].

Дискриминационная способность скоринговых функций означает, что оценка правильного PSM намного выше, и поэтому она хорошо отделена от распределения оценок неправильных PSM (т.е. нулевого распределения); следовательно, правильные PSM могут быть отделены от неправильных с помощью простого порога.
К сожалению, функции оценки при идентификации спектра затруднены (а) наличием множества необъяснимых пиков, которые возникают из-за необычной фрагментации пептида или загрязняющих молекул, или (б) отсутствием ожидаемых ионов фрагментации, которые не наблюдаются в масс-спектрометре [15].

14. What kind of score normalization (score calibration) techniques do you know?

Какие методы нормализации баллов (калибровки баллов) вы знаете?

Score normalization methods

• Analytical approaches (orange): binomial, hypergeometrial, normal distr.

• Linear approaches (purple): Comet's E-value, X!Tandem E-value

• Exact p-value (XPV) approaches (blue): MSGF+, Tide-search

• Heuristic (not based on statistical p-value)(Yellow): Xcorr, Tailor method

• Analytical approaches. They fit a distribution model to the empirical data:

• Linear approaches. They fit a distribution model to the empirical data X!Tandem and Comet use a log linear fit

• Exact approaches. An idea is that: we enumerate all possible peptide sequences, match them against an experimental spectrum database, and build the score histogram.

Методы нормализации баллов

- Аналитические подходы (оранжевый): биномиальный, гипергеометрический, нормальное распределение.

- Линейные подходы (фиолетовый): E-значение Комета, X!Tandem E-значение

- Подходы с точным p-значением (XPV) (синие): MSGF+, Tide-search

Эвристические (не основанные на статистическом p-значении) (желтые): Xcorr, метод Tailor,

Аналитические подходы. Они подгоняют модель распределения к эмпирическим данным:

- Линейные подходы. Они подгоняют модель распределения к эмпирическим данным. X!Tandem и Comet используют логарифмически линейную модель.

- Точные подходы. Идея заключается в следующем: мы перечисляем все возможные пептидные последовательности, сопоставляем их с экспериментальной базой данных спектров и строим оценку

их с экспериментальной базой данных спектров и строим гистограмму оценок гистограмма.

15. What is Peptide Prophet, how does it work? What are its benefits and limitations?

• PeptideProphet:
• Assumes there are two Gaussian hidden distributions:
• Distribution of incorrect scores
• Distribution of correct scores

• Uses Expectation-Maximization (EM) to fit these distributions

Uses these two Gaussians to calculate score statistics.
• Advantages:

• It can calculate a p- value of the hit, and it also can calculate FDR.
• Works over wide range of MS instruments, scoring functions, etc.

• Disadvantages:
• Assumes Gaussians, which is usually not the case.
• Assumes two components,
• There might by more components. (charge 2+, and charge 3+ peptides have different score distributions)
• In practice EM does not converge always, can stuck in local minima.
• Does not work with hyperscores (score of X!Tandem)
• In practice, histograms do not look as good as this plot

Peptide Prophet is a software tool.
Peptide Prophet works by analyzing the mass spectra of peptide ions and assigning probabilities to each peptide identification based on the quality of the fragmentation data. It uses statistical models to calculate these probabilities, taking into account factors such as the number of observed fragment ions, their intensity, and their consistency with the expected peptide sequence.

- PeptideProphet:
- Предполагает наличие двух гауссовских скрытых распределений:
- Распределение неправильных оценок
- Распределение правильных оценок

Использует метод Expectation-Maximization (EM) для подгонки этих распределений.

Использует эти два гауссиана для расчета статистики оценок.
- Преимущества:

- Может вычислять p-значение попадания, а также вычислять FDR.
Работает с широким диапазоном инструментов MS, скоринговых функций и т.д.

- Недостатки:
- Предполагает гауссианы, что обычно не так.
- Предполагает две компоненты,
- Компонентов может быть больше. (пептиды с зарядом 2+ и пептиды с зарядом 3+ имеют разные распределения баллов).
- На практике EM не всегда сходится, может застрять в локальных минимумах.
- Не работает с гиперкорами (оценка X!Tandem).
- На практике гистограммы выглядят не так хорошо, как этот график

Peptide Prophet - это программное обеспечение.
Peptide Prophet работает путем анализа масс-спектров пептидных ионов и присвоения вероятности идентификации каждого пептида на основе качества данных фрагментации. Для расчета этих вероятностей используются статистические модели, учитывающие такие факторы, как количество наблюдаемых фрагментных ионов, их интенсивность и соответствие ожидаемой последовательности пептида.

16. Explain Exact p-value method, how does it work? What are its benefits and limitations?

Exact p-value (XPV) calculation.

• An idea is that: we enumerate all possible peptide sequences, match them against an experimental spectrum database, and build the score histogram. This histogram can then be used to estimate a p-value empirically of a given query score sq as,

• p sq =#(s>sq)
#s
, this means that, the p-value is estimated by the number of the scores larger than the query score sq, divided by the total number of scores in the histogram.
• This approach is computationally exhaustive.

• An efficient way to calculate an empirical score distribution for XCORR scoring function. The main idea is that, instead of generate all spectrum brute force, we utilize sub-calculations stored in a dynamic programming table.

Расчет точного p-значения (XPV).

- Идея заключается в следующем: мы перечисляем все возможные пептидные последовательности, сопоставляем их
их с экспериментальной базой данных спектров и строим гистограмму оценок гистограмму. Эта гистограмма затем может быть использована для оценки p-значения эмпирическим путем для заданной оценки запроса sq как,

- p sq =#(s>sq)
#s
Это означает, что p-значение оценивается по количество оценок, превышающих оценку запроса sq, деленное на на общее количество оценок в гистограмме.
- Этот подход является исчерпывающим с вычислительной точки зрения.

- Эффективный способ вычисления эмпирического распределения баллов для скоринговой функции XCORR
скоринговой функции. Основная идея заключается в том, что вместо того, чтобы генерировать весь
спектр методом перебора, мы используем подвычисления, хранящиеся в таблице динамического таблице программирования.

16. Explain Exact p-value method, how does it work? What are its benefits and limitations?

Advantage of XPV:
• it gives indeed a very well calibrated p-value estimation for scores.
• Time: is polynomial.

The Exact p-value (XPV) method in mass spectrometry analysis is a statistical method used to determine the significance of peptide identifications. It works by comparing the observed mass-to-charge ratio (m/z) of a peptide with the theoretical m/z values of all possible peptides in a database. The XPV method calculates the probability of obtaining the observed m/z value or a more extreme value by chance, assuming that the peptide is not present in the sample.

The benefits of the XPV method in mass spectrometry analysis include its ability to provide accurate and reliable results, especially in complex mixtures of peptides. It also allows for the testing of non-normal distributions and can be used to control the false discovery rate (FDR) in peptide identifications.

However, the limitations of the XPV method include its computational complexity, which can make it time-consuming and difficult to perform in large datasets. It also assumes that the database is complete and accurate, which may not always be the case. Additionally, it may not be appropriate for some types of mass spectrometry data, such as data from ion mobility spectrometry or matrix-assisted laser desorption/ionization.

Метод точного p-значения (XPV) в масс-спектрометрическом анализе - это статистический метод, используемый для определения значимости пептидных идентификаций. Он работает путем сравнения наблюдаемого отношения массы к заряду (m/z) пептида с теоретическими значениями m/z всех возможных пептидов в базе данных. Метод XPV рассчитывает вероятность получения наблюдаемого значения m/z или более экстремального значения случайно, предполагая, что пептид не присутствует в образце.

Преимущества метода XPV в масс-спектрометрическом анализе включают его способность обеспечивать точные и надежные результаты, особенно в сложных смесях пептидов. Он также позволяет тестировать ненормальные распределения и может использоваться для контроля коэффициента ложных обнаружений (FDR) при идентификации пептидов.

Однако к ограничениям метода XPV относится его вычислительная сложность, что может сделать его трудоемким и трудным для выполнения в больших наборах данных. Он также предполагает, что база данных является полной и точной, что может быть не всегда так. Кроме того, он может не подходить для некоторых типов масс-спектрометрических данных, таких как данные спектрометрии ионной подвижности или матрично-ассистированной лазерной десорбции/ионизации.

17. What is Target-Decoy analysis? How does it work, what are the main assumptions it is based on?

Target-Decoy approach

• For every real peptide we randomly generate a “fake” peptide.
• Real peptide are called target peptides, fake peptides called decoy peptides. Decoy peptides are marked and merged with the target peptides.
• The main characteristics of target and decoy peptide should be similar, e.g. amino acid frequencies, peptide neutral mass distribution, etc.
• We assume that for every spectrum annotation, annotated with a decoy peptide, there is another incorrect spectrum annotation, annotated with a target peptide with a similar score.

Decoy generation strategies:
- Reversed: Keep the first and the last amino acid fixed and reverse the order of the internal amino acids. E.g.: PEPTIDE -> PDITPEE
- Shuffled: Keep the first and the last amino acid fixed and shuffle the internal amino acids: E.g.: PEPTIDE -> PTDIEPE
Difference is minor between the two strategy. However, reverse approach is deterministic making the test and database-search reproducible.

Sometimes a generated decoy peptide is identical to a target peptide. Some programs remove such decoy peptides from the search space, but this does not happen often.

We assume that: if we find a decoy peptide in the ranked output, we assume that there is also an incorrect match to a target peptide having similar score.

Подход "мишень-обманка

- Для каждого настоящего пептида мы случайным образом генерируем "фальшивый" пептид.
- Настоящие пептиды называются целевыми пептидами, а поддельные пептиды - приманками.
пептидами. Ложные пептиды помечаются и объединяются с целевыми пептидами.
пептидами.
- Основные характеристики целевого и ложного пептидов должны быть одинаковыми, например, частоты аминокислот, распределение нейтральных масс пептидов и т.д.
- Мы предполагаем, что для каждой аннотации спектра, аннотированной с помощью приманки пептидом, существует другая неправильная аннотация спектра, аннотированная с целевым пептидом с аналогичной оценкой.

Стратегии генерации приманки:
- Реверсивная: Сохраняйте первую и последнюю аминокислоту фиксированной и измените порядок
внутренних аминокислот. Например: ПЕПТИД -> PDITPEE
- Перемешанная: сохраняем фиксированными первую и последнюю аминокислоту и перетасовываем внутренние аминокислоты.
кислоты: Например: ПЕПТИД -> PTDIEPE
Разница между этими двумя стратегиями незначительна. Однако, обратный подход является
детерминированный, что делает тест и поиск в базе данных воспроизводимыми.

Иногда сгенерированный пептид-приманка идентичен целевому пептиду. Некоторые программы
удаляют такие пептиды-приманки из пространства поиска, но это случается нечасто.

Мы предполагаем, что: если мы находим пептид-приманку в ранжированном результате, мы предполагаем, что существует также неправильное совпадение с целевым пептидом, имеющим аналогичную оценку.

17. What is Target-Decoy analysis? How does it work, what are the main assumptions it is based on?

In order to control the FDR, we need to have an estimation on the number of the incorrect spectrum annotations. This can be estimated with using the so-called target-decoy-search strategy. This approach works as follows. Each peptide in the reference peptide dataset, that is associated to real, existing peptide molecule, is called the target peptide. For each target peptide, we generate another random peptide sequence called decoy peptide that does not exist in the reference peptide dataset. Therefore, the reference peptide dataset consists of two types of peptides: target and decoy peptides. The decoy peptides are often generated from target peptides via either (1) reversing the non-terminal amino acids, ot (2) shuffling the non-terminal amino acids. In order to obtain an unbiased FDR estimation with target - approach, the following criteria must meet:

1. We must ensure that the main characteristics of the set of target and decoy peptides are very similar, for instance, the amino acid frequen- cies, the distribution of the precursor ion mass, peptide length, should be the same among the candidate peptides.

2. We assume that for every spectrum annotation which is assigned to a decoy peptide there is another incorrect spectrum annotation assigned to a target peptide with roughly the same score.

3. The number of the target and decoy peptides must be equal, otherwise a correction factor should be employed in the FDR calculation.

4. It is assumed that incorrect spectrum annotations are equally likely to receive either target or decoy peptides.

5. The target and decoy peptides are distinct and independently gener- ated.

Для того чтобы контролировать FDR, нам необходимо иметь оценку количества неправильных аннотаций спектра. Это можно оценить с помощью так называемой стратегии поиска цели-обманки. Этот подход работает следующим образом. Каждый пептид в наборе данных эталонных пептидов, который связан с реальной, существующей пептидной молекулой, называется целевым пептидом. Для каждого целевого пептида мы генерируем другую случайную пептидную последовательность, называемую пептидом-приманкой, которая не существует в наборе данных эталонных пептидов. Таким образом, набор данных эталонных пептидов состоит из двух типов пептидов: пептидов-мишеней и пептидов-приманок. Пептиды-приманки часто генерируются из целевых пептидов путем (1) реверсирования внеконцевых аминокислот или (2) перестановки внеконцевых аминокислот. Чтобы получить несмещенную оценку FDR с помощью целевого подхода, должны выполняться следующие критерии:
Мы должны убедиться, что основные характеристики набора целевых и приманки пептидов очень похожи, например, частоты аминокислот, распределение масс ионов-предшественников, длина пептида должны быть одинаковыми среди пептидов-кандидатов.

Мы предполагаем, что на каждую аннотацию спектра, присвоенную пептиду-обманке, приходится другая неправильная аннотация спектра, присвоенная целевому пептиду с примерно такой же оценкой.

Количество целевых и ложных пептидов должно быть одинаковым, иначе при расчете FDR необходимо использовать поправочный коэффициент.

Предполагается, что неправильные спектральные аннотации с одинаковой вероятностью получают либо целевые, либо ложные пептиды.

Пептиды-мишени и пептиды-приманки отличаются и генерируются независимо друг от друга.

18. How is q-value defined?

• Q-value of a PSM is the smallest alpha level in the FDR control when the PSM is accepted.
• E.G. if a PSM with score sq has q-value = 0.01 that means: the PSM is trusted (accepted) if we control the FDR of experiment at alpha=0.01, but it would not be accepted if we controlled it at alpha=0.00999 or smaller.
• Q-values can be calculated using p-values or using TDA.

- Q-значение PSM - наименьший уровень альфа в контроле FDR, когда когда PSM принимается.
- Например, если PSM с оценкой sq имеет q-value = 0,01, это означает: PSM является доверяется (принимается), если мы контролируем FDR эксперимента на уровне альфа=0,01, но он не будет принят, если мы контролируем его при альфа=0.00999 или меньше.
- Q-значения могут быть рассчитаны с помощью p-значений или с помощью TDA.

The q-value of a spectrum annotation is defined as the smallest α level so that it is accepted at the α level of FDR. For instance, the q-value of a PSM is 0.005 then it is accepted at 0.5 % FDR level, but it is not accepted at, say, 0.50001 % FDR level. Note that the q-value of a PSM depends not only on the spectrum and its corresponding set of candidate peptides, but it also depends on other spectrum annotations too.

The pseudocode of q-values calculation algorithm is shown by Algorithm 2

The example of q-values calculation using Algorithm 2 is provided in Table 3.5.

A typical results obtained with database-searching is often reported by the number of accepted spectrum annotations as a function of the q-values.

q-значение спектральной аннотации определяется как наименьший уровень α, при котором она принимается на α-уровне FDR. Например, q-значение PSM равно 0,005, тогда оно принимается на уровне 0,5 % FDR, но не принимается, скажем, на уровне 0,50001 % FDR. Обратите внимание, что q-значение PSM зависит не только от спектра и соответствующего набора пептидов-кандидатов, но и от других аннотаций спектра.
Псевдокод алгоритма расчета q-значений представлен в Алгоритме 2.
Пример расчета q-значений с использованием алгоритма 2 приведен в таблице 3.5.
Типичные результаты, полученные при поиске по базе данных, часто отражают количество принятых аннотаций спектров как функцию q-значений.

Lecture 5
19. What is missed cleavage and how is it handled?

The missed cleavages refers to the situation when the enzyme did not cleave the protein molecule at the predicted cleavage site. This might happen due to the fact that the cleavage site is inaccessible for the enzyme. This situation can be handled easy by simulating the missed cleavages, usually up to 2-3 missed cleavage sites in the in silico protein digestion. This step usually increases the number of the peptide sequences in the reference dataset by the 2-3 times.

It is relatively easy to handle this type of error and many search programs deal with it.
• One can define the number of missed cleavages to consider during in silico protein digestion.
• Usually, 1-2 missed cleavages are considered.

Пропущенное расщепление относится к ситуации, когда фермент не расщепил
молекулу белка в предсказанном месте расщепления. Это может произойти из-за
того, что место расщепления недоступно для фермента. Эту ситуацию
можно легко разрешить, смоделировав пропущенные расщепления, обычно до 2-3
пропущенных участков расщепления в процессе in silico переваривания белка. Этот шаг обычно
увеличивает количество пептидных последовательностей в эталонном наборе данных в
2-3 раза.

Справиться с этим типом ошибки относительно просто, и многие поисковые программы
справляются с ней.
- Можно определить количество пропущенных расщеплений, которые необходимо учитывать во время in silico
переваривания белка.
- Обычно учитывается 1-2 пропущенных расщепления.

20. What is unexpected cleavage and how is it handled?

The unexpected cleavage refers to the situation when the peptide molecule breaks into two-or-more parts, and the terminal of the results sub-peptides do not correspond to expected cleavage site. In order to handle this situation and include the corresponding peptide sequences into the reference data set, one needs to generate all peptides resulted from imperfect enzymatic digestion. The number of the peptides in the reference dataset for various in silico peptide generation is shown in Table 3.2

Неожиданное расщепление относится к ситуации, когда молекула пептида распадается на две или более частей, и терминалы получившихся субпептидов не соответствуют ожидаемому месту расщепления. Для того чтобы справиться с этой ситуацией и включить соответствующие пептидные последовательности в набор эталонных данных, необходимо сгенерировать все пептиды, полученные в результате несовершенного ферментативного переваривания. Количество пептидов в эталонном наборе данных для различных in silico генераций пептидов показано в таблице 3.2

Unexpected cleavages

· When a peptide breaks into smaller parts:
· There are 3 types of cleavages:
· Tryptic (Both end of the peptide result from tryptic digestion (certainly, protein end and protein
start are not considered))
- Semi-tryptic (Only one end of the peptide results from tryptic digestion)
- Non-tryptic

Неожиданные расщепления

- Когда пептид расщепляется на более мелкие части:
- Существует 3 типа расщепления:
- Триптическое (оба конца пептида являются результатом триптического переваривания (конечно, конец и начало белка
начало не рассматриваются))
Полутриптический (Только один конец пептида является результатом триптического переваривания)
Нетриптический

21. What are post-translational modification (PTM), how does it affect the peptide-spectrum matching?

Modifications can occur to peptide molecule. The typical modifications are (a) post-translational modifications (PTMs) which regulate the protein activity and function in vivo by attaching a small molecule, such as phosphor to certain amino acids or (b) chemical modifications indicates a modification when a small atom or molecule is attached to certain amino acids, for instance, an oxygen atom can attached to the methionine amino acid during sample preparation, and finally (c) modifications may include amino acid mutations as well. For computational data analysis, these modifications can be handled and identified with the same algorithms, therefore we refer to them as modifications and it is commonly abbreviated as PTM. To generate modified peptide sequences is straightforward; however, the number of modified peptides can grow combinatorially. The Table 3.3 shows the number of the modified peptide sequences for various number of PTMs allowed to be present in the modified peptides at the same time. The peptide sequence is of length 11 and we assumed that all amino acid can be modified by 5 different PTMs.

В пептидной молекуле могут происходить модификации. Типичными модификациями являются (а) посттрансляционные модификации (ПТМ), которые регулируют активность и функцию белка in vivo путем присоединения небольшой молекулы, такой как люминофор, к определенным аминокислотам или (б) химические модификации - это модификация, когда небольшой атом или молекула присоединяется к определенным аминокислотам, например, атом кислорода может присоединиться к аминокислоте метионин во время подготовки образца, и, наконец, (в) модификации могут также включать мутации аминокислот. Для вычислительного анализа данных эти модификации могут быть обработаны и идентифицированы с помощью одних и тех же алгоритмов, поэтому мы называем их модификациями и обычно сокращенно PTM. Генерировать модифицированные пептидные последовательности просто, однако количество модифицированных пептидов может расти комбинаторно. В таблице 3.3 показано количество модифицированных пептидных последовательностей для различного числа PTM, которые могут одновременно присутствовать в модифицированных пептидах. Пептидная последовательность имеет длину 11, и мы предположили, что все аминокислоты могут быть модифицированы 5 различными PTM.

PTM alters the weight of amino acids and the peptide, and results peak shifts in the spectrum

22. How does the targeted PTM identification work?

Targeted PTM identification. In this approach, the experimenter has to guess few modifications which might be in the sample and specify this modifications individually. All major search engines, e.g. Crux, Sequest, Mascow, X! Tandem, Andromeda, supports this

Целевая идентификация PTM. При таком подходе экспериментатор должен предположить несколько модификаций, которые могут быть в образце, и указать эти модификации индивидуально. Все основные поисковые системы, например, Crux, Sequest, Mascow, X!Tandem, Andromeda, поддерживают этот подход.

Targeted: experimenter has to guess the modifications present in the sample.

Almost all search engine supports it.

- Experimenter needs to guess the PTMs in the sample.
- During theoretical peptide generation, the modifications have to be included to the theoretical peaks.

23. How does untargeted PTM identification work?

Untargeted PTM identification. This approach employs a large collection of known modifications (PTM DB) and it uses some heuristics method to find PTMs by avoiding the combinatorial explosion of by generating all possible modified peptide sequences.

Нецелевая идентификация PTM. Этот подход использует большую коллекцию известных модификаций (PTM DB) и использует некоторые эвристические методы для поиска PTM, избегая комбинаторного взрыва при генерации всех возможных модифицированных пептидных последовательностей.

Uses a big list of databases
• Search space is limited but can be very huge.
• if we allow 5 of the 10 most frequent modifications to occur in a peptide at the same type, the search space grows 3 orders of magnitude.
• The growth is more dramatic if instead of 10 types of modifications we wish to consider all of roughly 500 known types.

Использует большой список баз данных
- Пространство поиска ограничено, но может быть очень большим.
- Если мы допустим, что 5 из 10 наиболее частых модификаций происходят в пептиде при одного типа, пространство поиска увеличивается на 3 порядка.
- Рост будет еще более значительным, если вместо 10 типов модификаций мы захотим рассмотреть все из примерно 500 известных типов.

24. How does DeNovo PTM identification work?

De novo PTM identification. This approach does not use any known modifications as reference, but it tries to identify modification in the sample by searching for certain regular patters in the spectrum data.

Идентификация PTM De novo. Этот подход не использует какие-либо известные модификации в качестве эталона, а пытается идентифицировать модификацию в образце путем поиска определенных регулярных паттернов в спектре данные.

Lecture 6

25. How does the size of the search space impact on the statistical confidence (p-value) of the peptide-spectrum-match (PSM)?

Как размер пространства поиска влияет на статистическую достоверность (p-значение) пептидно-спектрального совпадения (PSM)

Large search space size reduces the significance of the annotation score.

- This is true for all types of searches which extends the search space.
- If search space is too big, you might end up identifying less peptides than with using smaller set of candidate peptide sequences.

Using a well-calibrated p-values, the distribution of the p-values of the random matches do not change. This is due to the statistical calibration. Using sidak correction, the corrected p-values remain “true” p-values. Therefore, the distribution of corrected p-values of the random matches remain uniform. However, the p-values of the correct peptide-spectrum matches will undergo a stronger correction making them ‘less significant’ and more difficult to distinguish them from the random matches. Thus, we will loose some correct peptide-spectrum-matches.

There are two major challenges when one generates too many peptides with either (1) too many modifications, (2) using non-standard digestion rules such as semi- or non-tryptic digestion, or (3) combining fasta files of too many taxa.
One of the main challenges is simply computational, i.e. the spectrum annotation procedure might need significantly more time to match every spectrum against thousands of millions of candidate peptides. As discussed in the previous section the number of semi- or non-tryptic peptides as well as the number of the modified peptides can explode combinatorially.
The other main challenge is that spectrum annotation undergoes a sort of multiple testing correction; thus, a high score may not end up being significant and in turn it results in fewer number of annotations at any level of FDR. We discuss this in details for similarity like scores (such as XCorr) and for p-values.

При использовании хорошо откалиброванных p-значений распределение p-значений случайных совпадений не меняется. Это происходит благодаря статистической калибровке. При использовании сидака
поправка, скорректированные p-значения остаются "истинными" p-значениями. Поэтому распределение скорректированных p-значений случайных совпадений остается равномерным. Однако p-значения правильных совпадений пептид-спектр подвергаются более сильной коррекции, что делает их "менее значимыми".
более сильной коррекции, что делает их "менее значимыми" и более трудными для отличия от случайных совпадений.
отличить их от случайных совпадений. Таким образом, мы потеряем некоторые правильные пептидно-спектральных совпадений.

Существуют две основные проблемы, когда генерируется слишком много пептидов либо (1) со слишком большим количеством модификаций, (2) при использовании нестандартных правил переваривания, таких как полу- или нетриптические переваривания, или (3) при объединении фаста-файлов слишком большого количества таксонов.
Одна из основных проблем - просто вычислительная, т.е. процедура аннотации спектров может потребовать значительно больше времени для сопоставления каждого спектра с тысячами или миллионами пептидов-кандидатов. Как обсуждалось в предыдущем разделе, число полу- или нетриптических пептидов, а также число модифицированных пептидов может вырасти комбинаторно.
Другая основная проблема заключается в том, что спектральная аннотация подвергается своего рода коррекции множественного тестирования; таким образом, высокий балл может оказаться не значимым, что в свою очередь приводит к уменьшению числа аннотаций при любом уровне FDR. Мы подробно обсуждаем это для оценок сходства (таких как XCorr) и для p-значений.

25. How does the size of the search space impact on the statistical confidence (p-value) of the peptide-spectrum-match (PSM)?

Как размер пространства поиска влияет на статистическую достоверность (p-значение) пептидно-спектрального совпадения (PSM)

The size of the search space can have a significant impact on the statistical confidence (p-value) of the peptide-spectrum-match (PSM) in mass spectrometry analysis. A larger search space means that there are more possible peptide sequences to consider, which can increase the likelihood of false positives and decrease the statistical significance of a PSM.

For example, if the search space includes only a few hundred peptides, then a PSM with a low p-value (e.g., $p < 0.05$) would be highly significant and unlikely to occur by chance. However, if the search space includes millions of peptides, then a PSM with the same p-value may not be as significant because there are many more possible peptides that could match the spectrum by chance.

To address this issue, researchers can use various methods to reduce the size of the search space, such as filtering out low-quality spectra or using more specific search parameters. This can improve the statistical significance of PSMs and reduce the false discovery rate (FDR).

Размер пространства поиска может оказать значительное влияние на статистическую достоверность (p-значение) пептид-спектрального совпадения (PSM) в масс-спектрометрическом анализе. Большее пространство поиска означает, что необходимо рассмотреть больше возможных пептидных последовательностей, что может увеличить вероятность ложноположительных результатов и снизить статистическую значимость PSM.

Например, если пространство поиска включает всего несколько сотен пептидов, то PSM с низким значением p-value (например, $p < 0,05$) будет высоко значимым и вряд ли произойдет случайно. Однако если пространство поиска включает миллионы пептидов, то PSM с таким же p-значением может быть не столь значимым, поскольку существует гораздо больше возможных пептидов, которые могут случайно совпасть со спектром.

Для решения этой проблемы исследователи могут использовать различные методы уменьшения размера пространства поиска, например, отфильтровывая низкокачественные спектры или используя более конкретные параметры поиска. Это может повысить статистическую значимость PSM и снизить коэффициент ложных обнаружений (FDR).

26. How does the size of the search space impact on the number of the annotations in the target-decoy search?

As a consequence, the sample-specific null distribution for the random matches shifts rigthward, but the score distribution of the correct annotations do not change. (Note that: the number of the correct peptide-spectrum-match might increase, but not the scores of the correct annotation. The expected number of matching peaks does not depend on the size of the search space)
Therefore the distributions of random matches and the correct matches close up. Since we usually control the false discovery rate, the decision threshold also needs to be adjusted and the threshold has to be increased. Since the distribution of the correct annotations does not change, we will loose some correct peptide-spectrum-matches.

The size of the search space can impact the number of annotations in the target-decoy search in mass spectrometry analysis. A larger search space means that there are more possible peptide sequences to consider, which can increase the number of annotations in the target-decoy search. This is because a larger search space increases the likelihood of finding false positives, which will be included in the decoy database and lead to more annotations.

On the other hand, reducing the size of the search space through filtering or more specific search parameters can decrease the number of annotations in the target-decoy search. This is because a smaller search space reduces the likelihood of finding false positives, resulting in fewer annotations in the decoy database and a lower false discovery rate (FDR).

Overall, the size of the search space is an important consideration in mass spectrometry analysis as it can impact the statistical significance of PSMs and the number of annotations in the target-decoy search.

Последствия при поиске цели-декоя

Как следствие, специфическое для выборки нулевое распределение для случайных совпадений смещается в сторону, но но распределение баллов правильных аннотаций не меняется. (Обратите внимание, что: количество правильных пептид-спектр-совпадений может увеличиться, но не количество баллов правильных аннотаций. Ожидаемое количество совпадающих пиков не зависит от размера пространства поиска).
Поэтому распределения случайных совпадений и правильных совпадений сближаются. Поскольку мы обычно контролируем коэффициент ложных обнаружений, порог принятия решения также должен быть скорректирован, а порог должен быть увеличен. Поскольку распределение правильных аннотаций не меняется, мы потеряем некоторые правильные пептидно-спектральные совпадения.

Размер пространства поиска может повлиять на количество аннотаций при поиске цели-разоблачителя в масс-спектрометрическом анализе. Большее пространство поиска означает, что существует больше возможных пептидных последовательностей, которые необходимо рассмотреть, что может увеличить количество аннотаций при поиске цели-обманки. Это происходит потому, что при большем пространстве поиска увеличивается вероятность обнаружения ложноположительных результатов, которые будут включены в базу данных приманки и приведут к увеличению числа аннотаций.

С другой стороны, уменьшение размера пространства поиска с помощью фильтрации или более конкретных параметров поиска может уменьшить количество аннотаций при поиске цели-приманки. Это объясняется тем, что меньшее пространство поиска снижает вероятность обнаружения ложных срабатываний, что приводит к меньшему количеству аннотаций в базе данных приманки и более низкому коэффициенту ложных обнаружений (FDR).

В целом, размер пространства поиска является важным фактором в масс-спектрометрическом анализе, поскольку он может влиять на статистическую значимость PSMs и количество аннотаций в поиске цели-приманки.

27. What is the Group FDR method and why was it developed?

The Group FDR method is a statistical method developed for mass spectrometry analysis to address the issue of multiple testing and false discovery rate (FDR) estimation. It was developed to improve the accuracy of FDR estimation by taking into account the correlation between peptide identifications.

In traditional FDR estimation methods, each peptide identification is considered independently, which can lead to an overestimation of the FDR when multiple peptides are derived from the same protein. The Group FDR method addresses this issue by grouping peptide identifications based on their shared protein identifications and estimating the FDR at the protein level.

By using the Group FDR method, researchers can more accurately estimate the FDR and reduce the number of false positive identifications in their mass spectrometry analysis. This can improve the reliability of downstream analyses and increase confidence in the results obtained.

Метод группового FDR - это статистический метод, разработанный для масс-спектрометрического анализа с целью решения проблемы множественного тестирования и оценки частоты ложных обнаружений (FDR). Он был разработан для повышения точности оценки FDR путем учета корреляции между пептидными идентификациями.

В традиционных методах оценки FDR каждая пептидная идентификация рассматривается независимо, что может привести к завышению FDR, когда несколько пептидов получены из одного и того же белка. Метод группового FDR решает эту проблему, группируя пептидные идентификации на основе их общих белковых идентификаций и оценивая FDR на уровне белка.

Используя метод группового FDR, исследователи могут более точно оценить FDR и уменьшить количество ложноположительных идентификаций в масс-спектрометрическом анализе. Это может повысить надежность последующих анализов и увеличить доверие к полученным результатам.

28. Why cannot you select only a certain type of spectrum annotation from the whole search results?

In mass spectrometry analysis, selecting only a certain type of spectrum annotation from the whole search results can lead to biased results and incomplete analysis. This is because different types of spectrum annotations provide different information about the identified peptides and proteins, and excluding certain annotations can result in missing important information.

For example, selecting only high-confidence peptide identifications based on a specific scoring algorithm may exclude valid identifications that do not meet the strict criteria. Similarly, excluding certain post-translational modification annotations can result in missing important information about protein function and regulation.

Therefore, it is important to consider all available spectrum annotations in mass spectrometry analysis to obtain a comprehensive understanding of the identified peptides and proteins. This can help researchers to draw accurate conclusions and make meaningful discoveries.

В масс-спектрометрическом анализе выбор только определенного типа аннотации спектра из всех результатов поиска может привести к необъективным результатам и неполному анализу. Это связано с тем, что различные типы спектральных аннотаций предоставляют разную информацию об идентифицированных пептидах и белках, и исключение некоторых аннотаций может привести к пропуску важной информации.

Например, отбор только высокодостоверных пептидных идентификаций на основе определенного алгоритма оценки может исключить достоверные идентификации, которые не соответствуют строгим критериям. Аналогично, исключение аннотаций определенных посттрансляционных модификаций может привести к упущению важной информации о функции и регуляции белка.

Поэтому важно учитывать все доступные спектральные аннотации при масс-спектрометрическом анализе, чтобы получить полное представление об идентифицированных пептидах и белках. Это может помочь исследователям сделать точные выводы и совершить значимые открытия.