Lily Neibart and Shannon Brady

Bio 110

November 9, 2014

Sequence alignment, Homology, and Paralogy

Outline:

**Bioinformatics**

* Conceptualizing biology in terms of molecules and applying “informatics techniques” to understand and organize the information associated with these molecules, on a large scale. Bioinformatics is a management information system for molecular biology and has many practical applications
  + Goals:
    - Giving access to existing information
    - Develop tools that aid in data analysis (BLAST)
    - Use these tools to analyze large sets of data and address questions in life science

**BLAST (Basic Local Alignment Search Tool)**

* By aligning mRNA with DNA we are able to find similar sequences in a database
  + How does is work?
    - Breaks up search sequence into short “words”
    - After finding matched it extends searched to upstream and downstream sequences
      * The query is the genetic information that we have
      * The subjects are located in the data base
      * The Blast provides the alignment of the query and the subject

Species Relationships Based on Anatomical Similarities

* Classify species based on shared features
* Anatomical features, such as presence of backbone, are used to study relationships

Species Relationships based on sequence similarities

* Molecular sequences (DNA, RNA, Protein) can be used for comparisons
* Anatomical comparisons looks for similarities and difference to deduce relationships
* Sequence comparisons can be more objective and less ambiguous than anatomical comparison

Sequence and Evolutionary Relationships

* Evolutionary related organisms share a common ancestor with an ancestral DNA sequence
* An organism evolved and diverges, their DNA sequences accumulate differences, also known as mutations
* Two common types of mutations are called SNPs and Indels
  + SNPs-type of mutation that can be used to track common ancestry in the single nucleotide polymorphism. A SNP is a change of one DNA base pair into another
  + Indels-another useful class of DNA mutation for tracing relatedness is insertions and deletions. These are mutations in which one of more nucleotide pairs in a DNA strand is lost of gained

DNA similarities and Evolutionary relationships

* Distantly related species have more time elapse since they shared a common ancestor; more time for mutations to accumulate in their DNA
* Closely related species have had little time to accumulate mutations since their last common ancestor; their DNA sequence are more similar

Aligning Sequences to Compare

* Before you can compare sequences, you need to make sure that you are comparing stretched of DNA that are evolutionary related (homologous sequences)
* Two sequences need to be aligned, so that related parts of sequences are lined up against each other and compared

Sequence Alignment: How does it work?

* Try to maximize the number of matched among all sequences being aligned by changing the positions of the sequences relative to one another and/or by adding gaps
* **E value**- The “expectations value”-number of alignments with this query sequences that would be expected to occur by chance with this database
  + The lower the E value the closer a fit of the nucleotide sequences
  + The similarities and differences between the query and the subject depends on insertion and deletion of sequences

**Homologous-**Characters in different organisms that is similar because they were inherited from a common ancestor

* Forelimbs in birds
  + Anatomical Homologies
    - The similarities between whale and hummingbirds is sue to their common ancestor

**Paralogy-** The same gene duplicated within the same species

* Hemoglobin alpha and beta are both found in humans

**Analogous**- have separate evolutionary organism, but are specifically similar because they have both experience natural selection. Two species evolve independently from a different ancestor, but display the same trait

* Bat and bird wings evolved independently by convergent evolution
* Wings of birds, bats and pterosaurs are convergent
  + Each group evolved independently from different wingless ancestors

**Anatomical Homologies**

* The humerous, radius, and ulna are a conserved feature that is found in all tetrapods
* Sometimes homologous features are no longer needed and they become vestigial
* Homologies are extended to the cellular level
  + Pant cells and animal cells have a lot in common
* Also extend on the molecular level
  + CO1 is found within species

5 Multiple Choice questions:

1. What is a paralogous structure? (a)
2. When the same gene is duplicated within the same species
3. When organisms evolve structures that are different from their evolutionary ancestors
4. When two unrelated species have a common trait

2. What does the E value represent? (b)

1. How similar paralogous structures are
2. How similar a query and a subject are
3. How similar an insert is to a deletion

3. What is an example of a analogous treat? (b)

1. Forelimbs
2. Wings
3. Beak
4. Color

4. What is an SNP? (a)

1. A change of one DNA base pair into another
2. A mutations in which one of more nucleotide pairs in a DNA strand is lost of gained
3. How similar a query and a subject are
4. When two unrelated species have a common trait

5. What are the goals of Bioinformatics? (c)

1. Develop tools that aid in data analysis (BLAST)
2. Use these tools to analyze large sets of data and address questions in life science
3. All the above

2 Essays:

Explain how gene alignment works.

In order to get a sequence of DNA, PCR and Sanger sequencing must be performed. Once the sequence is obtained it can be put into a Blast. Blast stands for Basic Local Alignment Search Tool. The sequence (which is the query) will be compared to subjects in the Blast database. The database will compare your query to different DNA sequences to see if they match any others. An E Value will be given with each match. The E value represents how similar or different the query is from the subject. The lower the E value the more similar the sequences are.

Explain the mutations that can occur during DNA sequencing.

During DNA sequencing the two common types of mutations are called SNPs and Indels. SNPs are a type of mutation that can be used to track common ancestry in the single nucleotide polymorphism. A SNP is a change of one DNA base pair into another. This is the position where the letters in a column do not match. An Indels is another useful class of DNA mutation for tracing relatedness is insertions and deletions. These are mutations in which one of more nucleotide pairs in a DNA strand is lost of gained. An indel in an alignment will be represented by a gap or sometimes a dash in the alignment.