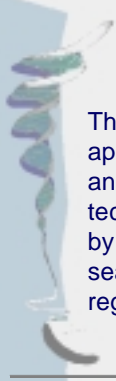


Introduction to Bioinformatics
<http://www.sdsc.edu/~helgew/bioinform/>

UCSD Extension, Fall 2000
Section 026380-5002/Course EE-40039
Helge Weissig, Ph.D.


Introduction to Bioinformatics • <http://www.sdsc.edu/~helgew/bioinform/> • © Helge Weissig, 2000 (1)



What is Bioinformatics?

The systematic development and application of computing systems and computational solution techniques analyzing data obtained by experiments, modeling, database search, and instrumentation regarding Biological aspect.


Introduction to Bioinformatics • <http://www.sdsc.edu/~helgew/bioinform/> • © Helge Weissig, 2000 (2)



What is Computational Biology?

Often used interchangeably with the term Bioinformatics. The systematic development and application of computing systems and computational solution techniques to models of biological phenomena;


Introduction to Bioinformatics • <http://www.sdsc.edu/~helgew/bioinform/> • © Helge Weissig, 2000 (3)



What's in it for me?

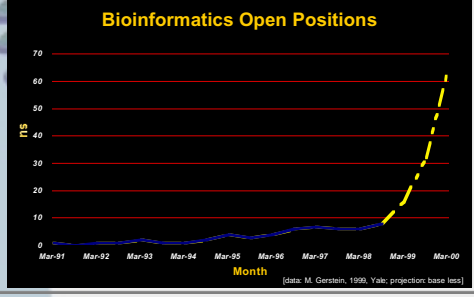
- o Course Goals:
 - o Provide overview and some theoretical insight into the most commonly used online Bioinformatics tools
 - o Practice query, retrieval and analysis of sequences and structures
 - o Gain familiarity with the terminology of computer/internet geeks

Introduction to Bioinformatics • <http://www.sdsc.edu/~helge/wbioinform/> • © Helge Weissig, 2000 (4)



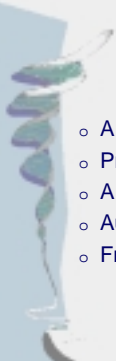
What's in it for me? [2]

Bioinformatics Open Positions



Month	N (Open Positions)
Mar-91	0
Mar-92	0
Mar-93	0
Mar-94	0
Mar-95	0
Mar-96	0
Mar-97	0
Mar-98	2
Mar-99	10
Mar-00	65

Introduction to Bioinformatics • <http://www.sdsc.edu/~helge/wbioinform/> • © Helge Weissig, 2000 (5)



What's NOT in it for me?

- o A degree in molecular biology
- o Programming knowledge & exercises
- o A guaranteed job offer after completion
- o Automatic fame and fortune
- o Free lunch

Introduction to Bioinformatics • <http://www.sdsc.edu/~helge/wbioinform/> • © Helge Weissig, 2000 (6)

Course Schedule

- 1st Class (today):**
 - Introduction and short student survey
 - More on Bioinformatics
 - Introduction to NCBI's Entrez
 - Hands-on: Entrez tutorial and query
 - Lunch
 - Introduction to BLAST
 - Hands-on: BLAST searches
 - Other Genetic Analysis tools
- 2nd Class (next week):**
 - Introduction to Protein Analysis
 - Hands-on session
 - Lunch
 - Introduction to Structural Bioinformatics
 - Hands-on session
 - "the other stuff":
 - Programming languages
 - Conferences
 - Online resources

Introduction to Bioinformatics • <http://www.sdsc.edu/~helgew/bioinform/> • © Helge Weissig, 2000 (7)

More on Bioinformatics

- General types:**
 - Databases (building, querying, object DB)
 - Text string comparisons (keyword search, alignments)
 - Pattern matching (AI/machine learning, clustering, data mining)
 - Geometry (robotics, visualization, docking)
 - Physical simulation (molecular mechanics, molecular dynamics)

Introduction to Bioinformatics • <http://www.sdsc.edu/~helgew/bioinform/> • © Helge Weissig, 2000 (8)

Components of Bioinformatics

Biological Data [S. Strabel's Genetics Lecture]


Computers [www.afcc.edu]

Algorithms

```
graph TD; A[Output] --> B{ }; B --> C[Input]; B --> D[ ]; D --> E[ ];
```

Introduction to Bioinformatics • <http://www.sdsc.edu/~helgew/bioinform/> • © Helge Weissig, 2000 (9)

What Data???



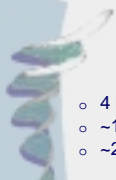
DNA → RNA → Protein → Phenotype
"central dogma of molecular biology"

Molecules:
Sequence, Structure, Function

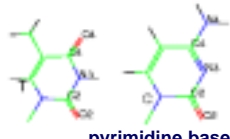
Processes:
Mechanism, Specificity, Regulation

Introduction to Bioinformatics • <http://www.sdsc.edu/~helge/wbioinform/> • © Helge Weissig, 2000 (10)

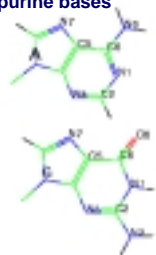
Sequence - DNA



- 4 bases: A, G, C, T
- ~1 k/gene
- ~2 M/genome




pyrimidine bases




purine bases

Introduction to Bioinformatics • <http://www.sdsc.edu/~helge/wbioinform/> • © Helge Weissig, 2000 (11)

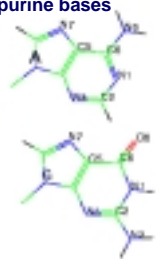
Sequence - RNA



- 4 bases: A, G, C, U
- very short to long



pyrimidine bases



purine bases

<http://www.embl-jena.de/>

Introduction to Bioinformatics • <http://www.sdsc.edu/~helge/wbioinform/> • © Helge Weissig, 2000 (12)

Sequence - Proteins

- 20 amino acids: [A-Z] but not [BJOUXZ]
- ~300 aa/protein (bacteria), ~200 aa/domain
- ~200 k known sequences

Introduction to Bioinformatics • <http://www.sdsc.edu/~helgew/bioinform/> • © Helge Weissig, 2000 (13)

Sequences - Information Growth

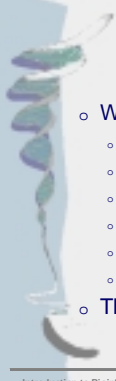
Introduction to Bioinformatics • <http://www.sdsc.edu/~helgew/bioinform/> • © Helge Weissig, 2000 (14)

Sequences - Genomes

1995 Bacteria, 1.6 Mb, ~ 1600 genes [Science 269: 496]	1997 Eukaryote, 13 Mb, ~ 6 k genes [Nature 387: 1]	1998 Animal, ~100 Mb, ~20 k genes [Science 282: 1945]	2000? Human, ~3 Gb, ~100 k genes [a patent office near you?]
--------------------------------------------------------------	----------------------------------------------------------	-------------------------------------------------------------	--------------------------------------------------------------------

© M. Gerstein, 1999, Yale (modified)

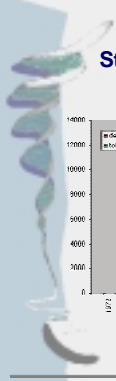
Introduction to Bioinformatics • <http://www.sdsc.edu/~helgew/bioinform/> • © Helge Weissig, 2000 (15)



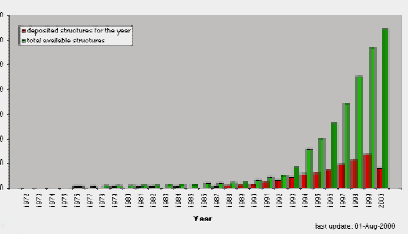
Sequences - Genomes [2]

- o Whole genome experiments:
 - o Micro arrays to analyze expression patterns
 - o Transposon & protein expression
 - o Systematic knock-outs
 - o 2 hybrids/linkage maps
 - o Metabolic pathways
 - o Regulatory networks
- o This is the beginning of a revolution!!

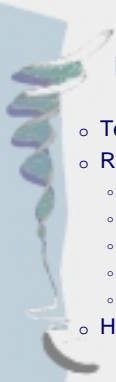
Introduction to Bioinformatics • <http://www.sdsc.edu/~helge/wbioinform/> • © Helge Weissig, 2000 (16)



Structural Information - Data Growth



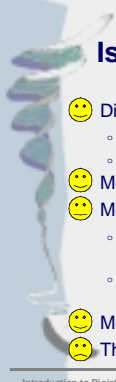
Introduction to Bioinformatics • <http://www.sdsc.edu/~helge/wbioinform/> • © Helge Weissig, 2000 (17)



Bioinformatics challenges

- o Telling signal from background
- o Redundancy and multiplicity:
 - o Different sequences with similar structures
 - o Organisms with similar genes
 - o Multiple functions of single genes
 - o Grouping of genes in pathways
 - o Sequence redundancy in genomes
- o How to find the similarities?

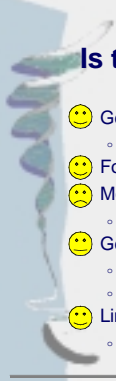
Introduction to Bioinformatics • <http://www.sdsc.edu/~helge/wbioinform/> • © Helge Weissig, 2000 (18)



Is this really Bioinformatics?

- 😊 Digital libraries
 - automated bibliographic search and textual comparison
 - knowledge bases for biological literature
- 😊 Motif discovery using sampling techniques
- 😊 Methods for structure determination
 - Computational crystallography
 - Refinement
 - NMR structure determination
 - Distance Geometry
- 😊 Metabolic pathway simulation
- 😊 The DNA computer

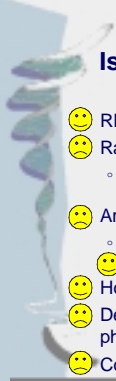
Introduction to Bioinformatics • <http://www.sdsc.edu/~helge/wbioinform/> • © Helge Weissig, 2000 (19)



Is this really Bioinformatics? [2]

- 😊 Gene identification via sequence analysis
 - Prediction of splice sites
- 😊 Forensic DNA methods
- 😊 Modeling of populations of organisms
 - Ecological modeling
- 😊 Genome sequencing methods
 - Assembling contigs
 - Physical and genetical mapping
- 😊 Linkage analysis
 - Linking specific genes to various traits


Introduction to Bioinformatics • <http://www.sdsc.edu/~helge/wbioinform/> • © Helge Weissig, 2000 (20)



Is this really Bioinformatics? [3]

- 😊 RNA structure prediction
- 😊 Radiological image processing
 - Computational representation of human anatomy ("visible human")
- 😊 Artificial life simulations
 - Artificial immunology/computer security
- 😊 Genetic algorithms in molecular biology
- 😊 Homology modeling
- 😊 Determination of phylogenies based on phenotypes
- 😊 Computerized diagnosis based on pedigrees

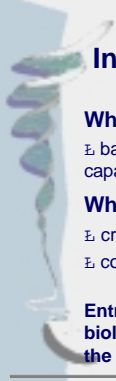
Introduction to Bioinformatics • <http://www.sdsc.edu/~helge/wbioinform/> • © Helge Weissig, 2000 (21)



Major Application Areas

- o Drug design
 - o Protein/ligand interaction studies
 - o Designing inhibitors
 - o Docking, structural modeling
- o Finding homologues
 - o Experiments in yeast are easier than in mice which are easier than in humans
- o Overall genome characterization
 - o Functional annotation!

Introduction to Bioinformatics • <http://www.sdsc.edu/~helge/wbioinform/> • © Helge Weissig, 2000 (22)



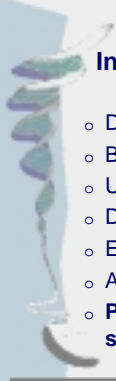
Introduction to NCBI's Entrez

What is Entrez???
E basic outline of the Entrez database system, its capabilities and some of its advanced features.

Why is Entrez important???
E cross-disciplinary reference database
E convenient query and retrieval capabilities

Entrez allows the retrieval of molecular biology data and bibliographic citations from the NCBI's integrated databases.


Introduction to Bioinformatics • <http://www.sdsc.edu/~helge/wbioinform/> • © Helge Weissig, 2000 (23)



Introduction to NCBI's Entrez [2]

- o Database structure
- o Basic Searching
- o Using neighbors
- o Details, Limits, Preview et al.
- o Entrez field specifiers
- o Advanced boolean searches
- o **Practical examples and hands-on session!**

Introduction to Bioinformatics • <http://www.sdsc.edu/~helge/wbioinform/> • © Helge Weissig, 2000 (24)

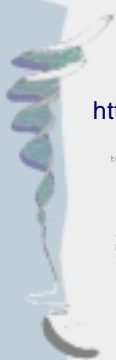


Database Structure

NCBI's integrated databases include:

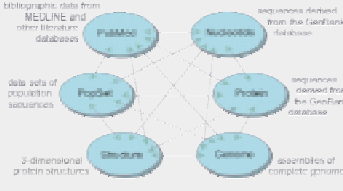
- o DNA sequences from GenBank, EMBL, and DDBJ
- o Protein sequences from Swiss-Prot, PIR, PRF, PDB, and translated protein sequences from the DNA sequence databases
- o Genome and chromosome mapping data
- o Three-dimensional protein structures derived from PDB incorporated into the Molecular Modeling Database (MMDB)
- o PubMed bibliographic database containing citations from the National Library of Medicine's MEDLINE and pre-MEDLINE databases
- o Aligned sequences as results from population, phylogenetic or mutation study to describe evolution and population variation (PopSet)

Introduction to Bioinformatics • <http://www.sdsc.edu/~helge/bioinform/> • © Helge Weissig, 2000 (25)




Database Structure [2]


<http://www.ncbi.nlm.nih.gov/Entrez/>



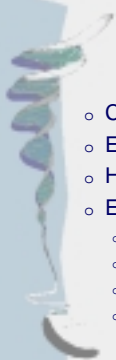
Introduction to Bioinformatics • <http://www.sdsc.edu/~helge/bioinform/> • © Helge Weissig, 2000 (26)



Database Structure [3]



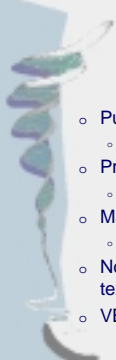
Introduction to Bioinformatics • <http://www.sdsc.edu/~helge/bioinform/> • © Helge Weissig, 2000 (27)



Basic Searching

- o Choose database
- o Enter query term
- o Hit "Go" button
- o Examples:
 - o **camp activated protein kinase**
 - o "camp activated protein kinase"
 - o "camp activated" "protein kinase"
 - o "infection*"

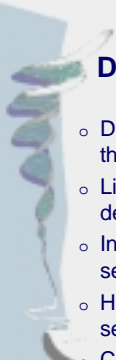
Introduction to Bioinformatics • <http://www.sdsc.edu/~helge/wbioinform/> • © Helge Weissig, 2000 (28)



Using neighbors

- o PubMed Neighbors
 - o Text & MeSH
- o Protein and Nucleotide neighbors
 - o BLAST
- o Macromolecular Structure Neighbors
 - o VAST
- o Non similarity based: references within text/annotation
- o VERY helpful in identifying related records

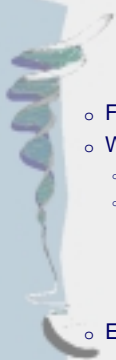
Introduction to Bioinformatics • <http://www.sdsc.edu/~helge/wbioinform/> • © Helge Weissig, 2000 (29)



Details, Limits, Preview et al.

- o Details: review Entrez's interpretation of the query term
- o Limits: ways to restrict a search to a defined subset of the database.
- o Indexes: alphabetical lists of terms from searchable database fields
- o History: review, recall and combine past searches
- o Clipboard: temporarily store results


Introduction to Bioinformatics • <http://www.sdsc.edu/~helge/wbioinform/> • © Helge Weissig, 2000 (30)



Entrez qualifiers

- o Format: **term [qualifier]**
- o Where:
 - o **Term** is the string to search for
 - o **Qualifier** specify the field to search in
 - o DB specific
 - o one of WORD, TITL, MESH, MAJR, AUTH, JOUR, ECNO, GENE, DATE, PDAT, MDAT, PAGE, VOL, KYWD, ORGN, ACCN, PROT, SLEN, SQID, SUBS, PROP, FKEY, and PTYP
- o Example: **Bourne PE [AUTH]**

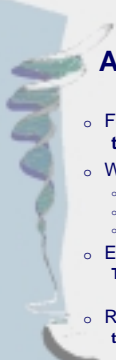
Introduction to Bioinformatics • <http://www.sdsc.edu/~helge/wbioinform/> • © Helge Weissig, 2000 (31)



Entrez qualifiers [2]

ACCN	accession number	AUTH	author name
DATE	publication year	ECNO	EC number
FKEY	feature key	GENE	gene name
JOUR	journal name	KYWD	keyword
MAJR	MeSH major topic	MDAT	modification date
PAGE	first page	ORGN	organism
PROP	property	PDAT	publication/creation date
PTYP	publication type	PROT	protein name
TITL	title word	SQID	sequence id
SLEN	sequence length	SUBS	substance
WORD	text word	VOL	volume

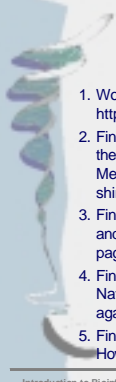
Introduction to Bioinformatics • <http://www.sdsc.edu/~helge/wbioinform/> • © Helge Weissig, 2000 (32)



Advanced boolean searches

- o Format: **term [field] operator term [field] ...**
- o Where **Operator** is any of :
 - o **AND** (intersection)
 - o **OR** (union)
 - o **BUT NOT** (difference)
- o Example: **Taylor P [AUTH] AND "protein kinases" [MESH]**
- o Ranges: **term1:term2 [field]**

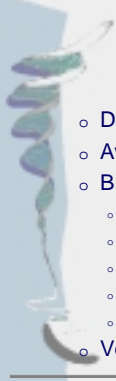
Introduction to Bioinformatics • <http://www.sdsc.edu/~helge/wbioinform/> • © Helge Weissig, 2000 (33)



Exercises

1. Work through the Entrez tutorial at <http://www.ncbi.nlm.nih.gov/80/Database/tut1.html>
2. Find references about shingles and facial paralysis. Display the records in the format that shows the abstract and the MeSH headings. How does PubMed map the term, shingles?
3. Find human nucleic acid sequences involved in apoptosis and cancer. Display all of the retrieved records on one Web page.
4. Find protein sequences related to AIDS with publications in Nature. Save this search strategy so the search can be run again at a later date.
5. Find all mouse or human genes added to Entrez in 1997. How many structures correspond to your result set?

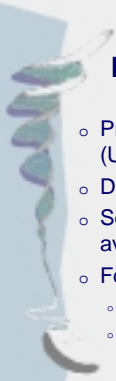
Introduction to Bioinformatics • <http://www.sdsc.edu/~helge/wbioinform/> • © Helge Weissig, 2000 (34)



Introduction to BLAST

- o Databases & data formats
- o Available search tools
- o BLAST
 - o The BLAST family of programs
 - o BLAST algorithm
 - o Result interpretation
 - o Advanced parameters
 - o Buyer beware!
- o VecScreen

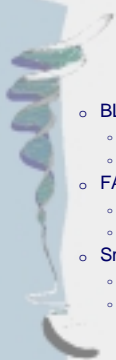
Introduction to Bioinformatics • <http://www.sdsc.edu/~helge/wbioinform/> • © Helge Weissig, 2000 (35)



Databases & Data Formats

- o Primary sequence repositories: GenBank (USA), EMBL (Europe) & DDBJ (Japan)
- o Daily data exchange and mirroring
- o Sequences submitted to one become available in all
- o Formats:
 - o FASTA format
 - o Complete database record

Introduction to Bioinformatics • <http://www.sdsc.edu/~helge/wbioinform/> • © Helge Weissig, 2000 (36)



Available search tools

- o BLAST (Basic Local Sequence Alignment Tool)
 - o Available on internet and downloadable
 - o Quick and simple
- o FASTA
 - o Available on internet and downloadable
 - o Primarily used in Europe
- o Smith-Waterman
 - o Very sensitive and highly accurate
 - o Very CPU intensive (usually requires hardware accelerators)


Introduction to Bioinformatics • <http://www.sdsc.edu/~helgew/bioinform/> • © Helge Weissig, 2000 (37)



Available search tools [2]

- o Which algorithm is best??
- o Depends on goal!
- o SW is most accurate and capable of detecting very weak similarities
- o Most similarity searches: BLAST, BLAST2.0
- o Fast and very accurate
- o Not designed for motif searching

Introduction to Bioinformatics • <http://www.sdsc.edu/~helgew/bioinform/> • © Helge Weissig, 2000 (38)




The BLAST family

Program	Query Sequence	Database Sequence Target
BLASTN	Nucleotide (both strands)	Nucleotide database
BLASTX	Nucleotide translated into 6 frames	Protein database
TBLASTX	Nucleotide translated into 6 frames	Nucleotide database translated in 6 frames
BLASTP	Protein	Protein database
TBLASTN	Protein	Nucleotide database translated in 6 frames

For searching proteins against nucleotide (including EST) sequences.

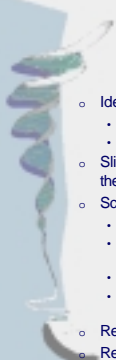
Introduction to Bioinformatics • <http://www.sdsc.edu/~helgew/bioinform/> • © Helge Weissig, 2000 (39)



GenBank Databases

Database	Sequence Type	Comments
NR	Nucleotide Protein	Contains "non-redundant" nucleic acid and protein sequences from all known data sources.
MONTH	Nucleotide	Contains new records added to NR within the last month.
SWISSPROT	Protein	The SWISS-PROT annotated subset of NR
DBEST	Nucleotide from cDNA source	Database of Expressed Sequence Tags
DBSTS	Nucleotide from genomic source	Database of Sequence Tagged Sites
PDB	Protein	The Brookhaven Protein Database subset of NR
VECTOR	Nucleotide	A database of vector sequences. Hasn't been updated in a long time. Useful to check sequences against before submitting them to GenBank.
KABAT	Protein	Proteins of immunological interest. A subset of NR
MITO	Nucleotide	A database of mitochondrial DNA sequences. Also useful to check before submitting sequence to GenBank.
ALU	Nucleotide	Database of repetitive DNA sequences. Check before submitting sequence to GenBank.
EPD	Nucleotide	Eukaryotic Promoter Database
YEAST	Protein	S. cerevisiae protein database
E. COLI	Protein	E. coli genomic coding region translations
HTGS	Nucleotide	High-Throughput genomic sequence. Usually single pass data

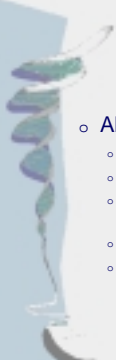
Introduction to Bioinformatics • <http://www.sdsc.edu/~helge/wbioinform/> • © Helge Weissig, 2000 (40)



The BLAST algorithm

- o Identify HSP's (High-scoring Segment Pairs)
 - default 11 bp or 3 aa
 - perfect match
- o Slide query and target sequence across each other until the maximum number of HSP for that target is found
- o Score the alignment
 - a scoring matrix is used (such as BLOSUM 62)
 - gaps introduced between HSP's during sliding get negative score
 - a match gets a positive score
 - total alignment score is subjected to statistical analysis to calculate the significance vs. chance of the score
- o Repeat for every sequence in the target database
- o Return total results


Introduction to Bioinformatics • <http://www.sdsc.edu/~helge/wbioinform/> • © Helge Weissig, 2000 (41)



Buyer Beware!

- o Alu repeats
 - o Most common repetitive DNA element
 - o 282 bp, ~ every 3300 bp in human DNA
 - o Several subfamilies with some sequence homology
 - o Just one example!
 - o Identifying:
 - o query with BLASTN against Alu db
 - o Query with BLASTX against SWISS-PROT
 - o contains dummy translations of Alu
 - o Hits are designated as Alu

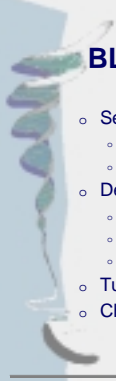
Introduction to Bioinformatics • <http://www.sdsc.edu/~helge/wbioinform/> • © Helge Weissig, 2000 (42)



Buyer Beware! [2]

- o Vector Sequences
 - o Artifacts from sequencing procedure
 - o Make sure to remove them in query sequence!
 - o Identifying:
 - o query with BLASTN against Vector db
- o EST data
 - o Single pass sequencing with up to 5% errors
 - o Many ESTs tag the same gene
 - o dbEST is best searched with BLASTX and TBLASTX
 - o dbEST is not broken down by species

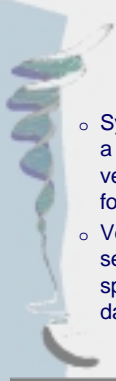
Introduction to Bioinformatics • <http://www.sdsc.edu/~helge/w/bioinform/> • © Helge Weissig, 2000 (43)



BLAST With Small Sequences

- o Set E-value to ≥ 1000
 - o Smaller sequences are more likely to occur by chance
 - o Increasing E results in more matches
- o Decrease word size (W)
 - o BLASTN needs $W \geq 7$
 - o Query length $\geq 2W$
 - o The smaller W, the slower the search
- o Turn filter option "OFF"
- o Change the scoring matrix

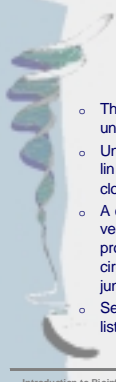
Introduction to Bioinformatics • <http://www.sdsc.edu/~helge/w/bioinform/> • © Helge Weissig, 2000 (44)



VecScreen

- o System for quickly identifying segments of a nucleic acid sequence that may be of vector origin using optimized parameters for blastn.
- o VecScreen searches a query for segments that match any sequence in a specialized non-redundant vector database (UniVec).

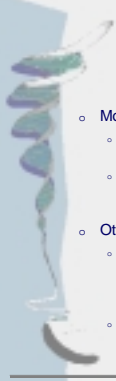
Introduction to Bioinformatics • <http://www.sdsc.edu/~helge/w/bioinform/> • © Helge Weissig, 2000 (45)



VecScreen Details

- The UniVec database contains only one copy of every unique sequence segment from a large number of vectors.
- UniVec also contains sequences for those adapters, linkers and primers commonly used in the process of cloning cDNA or genomic DNA.
- A copy of the first 49 bases of the sequence for a circular vector is appended to the end of the sequence before it is processed for addition to UniVec. This "pseudo-circularization" allows matches that span the circular junction to be identified correctly.
- Sequences used to build the current version of UniVec are listed in the UniVec Representation List.

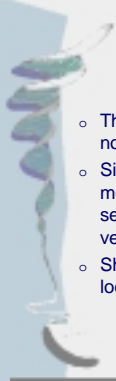
Introduction to Bioinformatics • <http://www.sdsc.edu/~helgew/bioinform/> • © Helge Weissig, 2000 (46)



Why use VecScreen?

- Most common sources of contamination:
 - Vectors -- Failure to identify and remove all the vector sequence results in a finished sequence that is contaminated.
 - Adapters, linkers, and PCR primers -- Often present in raw sequences and will contaminate the finished sequence unless they are identified and removed.
- Other sources of contamination:
 - Transposons and Insertion Sequences -- A transposable element from the cloning host (generally *Escherichia coli* or yeast) occasionally will insert itself into the cloned DNA/RNA while the clone is being propagated.
 - Impurities in the DNA/RNA under investigation -- Often derived from impure reagents, incomplete isolations or heterogeneous organism (fungal, bacterial) contaminants.

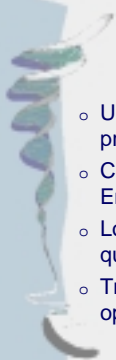
Introduction to Bioinformatics • <http://www.sdsc.edu/~helgew/bioinform/> • © Helge Weissig, 2000 (47)



Limitations of VecScreen

- The UniVec database was constructed such that no sequence element is longer than 50 bp.
- Since the database is made of fragments and most vectors contain many common regions, search results will not indicate the identity of the vector with the strongest match to the query.
- Should not be used in a case where you are looking for a match of longer than 50 bp.

Introduction to Bioinformatics • <http://www.sdsc.edu/~helgew/bioinform/> • © Helge Weissig, 2000 (48)



Exercise

- o Use Entrez to identify sequences (both protein and NA) for use with BLAST
- o Compare sequence neighbor results of Entrez with your results
- o Look at different E values for the same query sequence
- o Try out the ungapped and un-filtered options

Introduction to Bioinformatics • <http://www.sdsc.edu/~helge/bioinform/> • © Helge Weissig, 2000 (49)
