SENECA COLLEGE

BIF701 FINAL REPORT

A Bioinformatics Tool Safari

Author:
Christopher EELES

Supervisor:
Monica WONG

An exploration of the resources and tools available in the field of bioinformatics for the Bioinformatics Graduate Certificate

in the

School of Biological Sciences and Applied Chemistry

December 7, 2018

Contents

"This paper was created with the intention of expanding the scope as the author accumulates more knowledge and expertise in the field. Therefore many sections may be short or incomplete, with the possibility of expansion at a later date."

Author's Note

Chapter 1

Sequence Alignment

1.1 Concepts Discussed

For a detailed outline of BLAST concepts and operation please see Having a BLAST with bioinformatics.

1.1.1 PAM Matrices

See lecture notes.[1]

1.1.2 BLOSSUM

See lecture notes.[1]

1.2 Tools Discussed

1.2.1 BLAST

Basic Local Alignment Search Tool (BLAST) is a sequence analysis service provided by the National Centre for Biotechnology Information in the United States. A BLAST query is composed of four parts:

1. A query - what you want to find.
2. A database - where you are going to find it.
3. A program - how you are going to find it.
4. A purpose/goal - why you are going to find it.

A query is the set of information you wish to input into the alignment, it is usually in the form of an ID gathered from an online biological database.[2] Many databases are available which fall into two broad categories: protein and nucleotide.[2] Each category of databases contain specialized forms of information which should be considered when selecting which to use for a specific investigation.[2]

Program selection is also necessary to ensure the correct algorithms and parameters are used in your analysis. While many programs exist they fall into two major categories: nucleotide BLAST and protein BLAST.[3] Within each category there are several tools, including ones to translate nucleotide data into proteins and to reverse translate protein sequences into nucleotides.[3] An overview of available programs is provided below.

* * *

1.2.1.1 BLASTn

This program compares a query sequence against a nucleotide sequence database.[3] Within BLASTn there are two special programs allowing an interaction with the BLASTp program:

1. **BLASTx**
   This tool compares a nucleotide query sequence translated in all reading frames against a protein sequence database.[3]

2. **tBLASTx**
   This tool compares a six frame translation of the nucleotide query sequence against the six-frame translations of nucleotide sequences in a database.[3]

1.2.1.2 BLASTp

Similarly to BLASTn, this program compares an amino acid sequence query against a protein sequence database.[3] It also contains a specialized tool to interact with BLASTn:

1. **tBLASTn**
   This tool compares a protein sequence query against a nucleotide database dynamically translated in all reading frames.[3]

* * *

With this knowledge about databases, queries and programs one should be able to select combinations which meet their specific research interests.[2] BLAST is a powerful tool with access to vast quantities of sequence data and therefore should be one of the first tools considered for sequence alignment analyses.

1.3 Tool Usage

BLAST can be found at NCBI BLAST. The help file for BLAST can be found BLAST Help. Please reference the help file for instructions, troubleshooting and additional information as needed.

1.3.1 BLASTp

1. Navigate to BLASTp from the BLAST main page.

2. In the query sequence section enter accession number(s), GI number(s), or FASTA sequence(s) into the text box. Alternatively you may upload files from your local disk.

3. Select query sequence parameters such as query subrange, job title, or multiple alignments as needed.

4. In the choose search set section, select the appropriate database and choose optional parameters such as organism of interest, exclusions and entrez queries as needed.

5. In the program selection box choose the desired program. Specialized BLASTp algorithms are available which use of alternative scoring methods. These include PSI-BLAST, PHI-BLAST, etc.

6. Expanding the algorithm parameters options below the BLAST icon allows parameters to be select for the chosen algorithm; these include general, scoring parameters as well as filtering/masking conditions. This gives you a high degree of control over your BLASTp query and significantly expands utility for advanced users of the BLAST suit of web-apps.

7. The results page displays a wealth of information, including visual summaries of the top 100 alignments, description summaries of each alignment, as well as detailed alignment data.

1.4 Tool Applications

The applications of BLAST are to diverse to list exhaustively, but new entries will be added when encountered.

1.4.1 BLASTp

Identifying Protein Virulence Factors

This tool can be used to align novel gene sequences for proteins in one species of pathogen for comparison to another. The alignment allows identification and correlation of gene homologs in other species. If a gene is present in many pathogenic species, it may warrant closer investigation to confirm its role as a virulence factor.[4]

Chapter 2

Dynamic Programming

2.1 Concepts Discussed

2.1.1 Needleman-Wunsch Algorithm

See lecture notes.[5]

2.1.2 Smith-Waterman Algorithm

See lecture notes.[5]

2.2 Tools Discussed

2.2.1 Microsoft Excel

Microsoft Excel is a member of the Microsoft Office software suit used for generating, storing and manipulating tabular data.[6] While the features of Excel are quite limiting in terms of data analysis and visualization, it can be useful as an introductory tool for data entry, management, manipulation, analysis and visualization. The user friendly GUI allows new users to quickly understand and utilize the softwares functionality, thus enabling non-technical users to become acquainted with tabular data before being exposed to more complex analysis software such as R.[6] Despite these disadvantages Excel is still a common choice for data analysis in the scientific community, especially those in fields largely unrelated to computers or computation.

2.3 Tool Usages

2.3.1 Microsoft Excel

Excel can be used to input and store tabular data, perform basic and complex formulas as well as output various graphs and figures to visualize data.[7] Input and output is relatively simple, using drop down menus with clearly labeled commands; usefully, .csv files can be both loaded and exported using the GUI.[7] Basic formulas are included with the package for arithmetic and simple statistical calculations like the mean. More complex formulas can be built into cells using the Fx box just below the program ribbon. These formulas can be cell specific, as well as take in values from other cells to built a series calculations into a single Excel file.[7] You can also define your own formulas using the GUI for commonly used calculations. Essentially the program offers some of the functionality of a programming language without the need to learn the more complex concepts and syntax required to use more powerful tools.

2.4 Tool Applications

2.4.1 Microsoft Excel

2.4.1.1 Calculating Substitution Matrices

Substitution matrices involve a complex set of algorithmic instructions to output the score for each potential alignment.[5] Excel can be used as a simple way to dynamically program the correct substitution matrix for a given alignment using relatively simple functions in the program. Applying formulas to the values in an alignment allows input of relatively few values while outputting the entire matrix; completing this task by hand is tedious and time consuming, so even crude dynamic programming implementation have value.[8] Through the use of cell references, we can set absolute cell values for the initial values of the rows and columns.[8] Subsequently mixed cell references can be used to call the absolute values in adjacent cells.[5] Finally a network of relative cell references can be applied to the following cells in order to generate the matrix for a desired alignment.[5] This technique can be adapted for various matrices such as Needleman-Wunsch, Smith-Waterman, PAM and BLOSSUM allowing automated calculation for a given alignment.

Chapter 3

Database Searching

3.1 Concepts Discussed

3.1.1 Heuristic Algorithms

See lecture notes.[9]

3.1.2 Types of Databases

See lecture notes.[9]

3.1.3 Classification of Databases

See lecture notes.[9]

3.2 Tools Discussed

3.2.1 Primary Databases

See lecture notes.[9]

3.2.2 Metadatabases

3.2.2.1 Online Mendelian Inheritance in Man Database

OMIM is designed to be a knowledge-base for physicians and clinical genetics researchers which is comprehensive, authoritative, and timely.[10] The system was built to support human genetics research, eduction, and clinical genetics practice.[10] Curated by Johns Hopkins University Medical School and updated daily, the web-service provides a wealth of compiled information about clinically relevant phenotypes and their associated clinical manifestation as well as molecular and physiological manifestations.[10] It is accesible from any page of the NCBI Entrez suite, along with on the OMIM homepage. The service acts as a one-stop shop for collated information about clinically relevant genes.

3.2.2.2 Ensembl

A genome browser used to search across multiple databases.[9]

3.2.2.3 NCBI

A genome browser used to search across multiple databases.[9]

3.3 Tool Usages

3.3.1 OMIM Database

The OMIM web-page has a simple layout with only one text box to input a query. Accepted query types including MIM number, disorder name, gene name/symbol or plain English.[10] Entries in OMIM returned from a search are categorized into five symbols which represent whether they contain information on genes, phenotypes or both:[10]

1. * Indicates an entry containing a know sequenced gene.[10]
2. # Indicates an entry is descriptive, containing information about phenotype; these entries may not represent a unique genetic locus.[10]
3. + Indicates an entry contains a phenotype description along with information about a known gene.[10]
4. % Indicates an entry for which the description matches a confirmed Mendelian phenotype/locus where the underlying molecular mechanism is unknown.[10]
5. ^ Indicates an entry was removed or merged with another entry in the database.[10]
6. An entry without a symbol represent a gene with a suspected, but not confirmed Mendelian basis or that it may be a duplicate/redundant phenotype.[10]

The page returned from an OMIM query displays variable amounts of information, but high quality entries usually contain information about the gene to phenotype relationships as well as phenotypic and genotypic manifestations with through and linked references. Perhaps the most valuable feature of this database is the wealth of links provided in the right hand external links menu. For there one can access links to external sources which may provide more detailed descriptions of some facet of the MIM entry.[10]

3.4 Tool Applications

3.4.1 OMIM Database

3.4.1.1 Identifying Genes Associate with Disease

OMIM is a powerful search tool to determine what genes play a role in the development of genetic pathology. By searching the name of a disease, one is able to see associated genes; conversely, by searching a gene name, you can identify if it is known to cause genetic pathology.[11] From within a gene or disease entry in OMIM you can navigate to numerous external resources to allow rapid accumulation of specific information about the gene or phenotype.[10] The real utility for this database tool is to rapidly assemble information from diverse database entries for easy access on one web-page. This will be similar for other gene browsers, making them an essential tool in the bioinformaticians tool box.

Chapter 4

Multiple Sequence Alignment

4.1 Concepts Discussed

4.1.1 Position Specific Scoring Matrices (PSSMs)

See lecture notes.[12]

4.1.2 Position Specific Iterated (PSI) BLAST

See lecture notes.[12]

4.1.3 Cluster Alignment (CLUSTAL) Algorithms

See lecture notes.[12]

4.2 Tools Discussed

4.2.1 BlastN

The help file for BLAST can be found BLAST Help.

4.2.2 ClustalW

Information about ClustalW can be found at ClustalW2.

4.2.3 ClustalOmega

ClustalOmega is a web-based heuristic multiple sequence alignment tool used to evaluate similarity between biological sequence data.[13] The ClustalOmega system is scalable and widely viewed as one of the fastest online multiple sequence alignment tools; speed and accuracy for a small number of sequences are similar to other high quality sequence aligners.[13] However, the tool provides significance performance gains in comparison of large data sets with hundreds of thousands of sequences.[13] The ClustalOmega algorithm struck a balance between previous fast but error prone multiple sequence aligners and the next generation of highly accurate but computationally ones.[13] The program is able to provide reasonable accuracy without an excessive use of computational resources. Creation of a guide tree before alignment by vectorizing the distance between each aligned pair enables these performance gains.[13] Additional information from external sources can be added to an alignment to provide additional gains in accuracy during alignments.[13]

4.3 Tool Usage

4.3.1 ClustalOmega

The Clustal Omega multiple sequence alignment web-app allows for alignment of three or more sequences to produce biologically meaningful alignments for divergent sequences.[14] After alignment, detected evolutionary relationships can be visualized in phylogenic trees of various types.[14] Operation of ClustalOmega follows these steps:

1. The entry page allows input of multiple sequences in a variety of formats via upload or copy and paste to a text box; sequence can be RNA, DNA or protein.[14]

2. A number parameter settings are available via a drop down menu. Changing these values allows choice of alignment formatting.[14]

3. Advanced parameter options are revealed by the more options button. These allows choice of whether or not to include a guide tree, how many iterations of clustering should occur as well as number of guide tree iterations.[14]

4. The submission button is at the bottom of the page, with the option to be notified by email for longer compute times.[14]

5. The results page shows the outputs in the selected format with an $*$ under perfectly aligned positions.[14]

6. Additional results options include visualization and generation of phylgenetic trees, all of which can be selected from the tabs at the top of the results page.[14]

4.4 Tool Application

4.4.1 ClustalOmega

4.4.1.1 Identifying Gene Transfer Between Organisms

Clustal Omega provides a significant number of options for visualizing results. However, the multiple sequence alignment in itself can be useful for comparing genes or loci between or within species. By selecting as input the sequences suspected to be homologous and running the tool, the alignment data can be used to identify regions of similarity.[15] Such information can be used to hypothesize about whether sequences are likely to have similarly due to convergent evolution, or if some form of gene transfer—*i.e.,* plasmid exchanges, conjugation in bacteria; translocations in complex organisms—is a better explanation for the observed alignment.[15] ClustalOmega is a versatile tools for multiple sequence alignment, with additional functionality for fields such as phylogeny. Therefore this tool should be considered whenever a bioinformatician wants to explore genetic relationships between or within species.

Chapter 5

Evolutionary Processes

5.1 Concepts Discussed

5.1.1 Phylogenetic Trees

5.1.1.1 Gene Trees

See lecture notes.[16]

5.1.1.2 Species Trees

See lecture notes.[16]

5.1.1.3 Tree Roots

See lecture notes.[16]

5.1.1.4 Nodes

See lecture notes.[16]

5.1.1.5 Tree Topologies

See lecture notes.[16]

5.1.2 Clustering Methods

5.1.2.1 Unweighted Pair-Group Method Arithmetic Mean (UPGMA)

See lecture notes.[16]

5.1.2.2 Fitch-Margoliash Method

See lecture notes.[16]

5.1.2.3 Neighbour-Joining Method

See lecture notes.[16]

5.2 Tools Discussed

5.2.1 Phylogeny.fr

Considering the central role of phylogenetic analysis in many biological research areas, tools to automate such tasks are required.[17] Phylogeny.fr provides a transparent platform to chain together tools for identification and alignment of homologous sequences along with phylogenetic reconstruction and graphical representation of the inferred tree.[17] Phylogeny.fr provides an integrated platform for biologists unfamiliar with phylogenetic analysis to easily generate phylogenetic data and visualizations for use in their research.[17]This robust, ready-to-use pipeline uses MUSCLE for multiple alignment, PhyML for tree building, and TreeDyn for tree rendering all with parameters preset to suit most studies.[17] Inclusion of the "advanced" mode uses the same pipeline but allows the user to customize program parameters, while "a la carte" mode adds the ability to modify programs in the chain.[17] These features make phylogeny.fr a powerful resource for phyolgenetic analysis for both general biologists and phyogenetic specialists. It should, therefore, be near the top of the tool list for bioinformatics applications in phylogeny and adjacent fields.

5.2.2 Fneighbor

See documentation for fNeighbor.

5.3 Tool Usages

5.3.1 Phylogeny.fr

The phylogeny.fr program is available here. Choice of the level of customization must be chosen from the main page.

5.3.1.1 "One Click"

After selecting "One Click" from the homepage an input page loads. The analysis may be given a name, files uploaded from a local disk or pasted from clipboard in FASTA, EMBL, or NEXUS file format. Using the default settings the program can handle 200 sequences of maximum length 2000 for proteins or 6000 for nucleic acids. Options exist to apply "Gblocks" to eliminate poorly aligned positions as well as an email box should you wish to retrieve your analysis at a later date. Due to selection of "One Click" the settings tab is not used.

Results of the analysis populate the phylogeny.fr tabs starting at number three. The default page after analysis shows the sequence alignment results colourized based on the score of each position. This alignment data can be downloaded in FASTA, Phylip, and Clustal formats. Tab four displays a curated version of the alignment with poorly scoring positions removed; the percent of the original alignment is listed under outputs along with options to download the curated alignment in FASTA or Phylip formats. The phylogeny tab, number five, displays the phylogenetic tree inferred from the alignment data as text. Under outputs the substitution model, gama shape parameter, number of categories and proportion of the sequence which is invariant are listed; download link are provided for the tree in Newick format as well as the statistics file. The final tab, number six, creates a graphical rendering of the tree, allowing customization of the colours, labels and style of tree

displayed, with a wide range of options to customize your results to highlight relevant information. Images of the customized tree are available for download in PNG, PDF, SVG, TGF, Newick and text formats.

5.3.1.2 "Advanced"

5.3.1.3 "A la Carte'

5.3.2 Fneighbor

See documentation for fNeighbor.

5.4 Tool Applications

5.4.1 Phylogeny.fr

5.4.1.1 Determining The Degree of Relatedness Between Species on the Tree of Life

FASTA files from different families and domains on the tree of life can be input to phylogeny.fr to yield alignment data and a phylogenetic tree of the degree of inter-species relation.[18] This is based on the degree of genetic similarly in the alignments from the first program in the chain. The generated tree can be used to confirm suspected relations between species, or to infer the correct taxonomy for newly discovered species. Numeric data generated during the analysis can also be used to quantify the evolutionary distance between two species. Subsequent statistical analysis would allow inferences to be made about the strength of such relationships

5.4.2 Fneighbor

See documentation for fNeighbor.

Chapter 6

Gene Prediction

6.1 Concepts Discussed

6.1.1 Open Reading Frames

See lecture notes.[19]

6.1.1.1 Prokaryotic ORFs

See lecture notes.[19]

6.1.2 Upstream Coding Sequence

See lecture notes.[19]

6.1.3 Genome Features

See lecture notes.[19]

6.1.3.1 Prokaryotes

See lecture notes.[19]

6.1.3.2 Eukaryotes

See lecture notes.[19]

6.1.4 Gene Finding Schemes

6.1.4.1 Types of Algorithms

See lecture notes.[19]

6.1.4.2 Confirming Predictions

See lecture notes.[19]

6.1.5 Gene Ontology

See lecture notes.[19]

6.1.5.1 Syntenic Regions

See lecture notes.[19]

6.2 Tools Discussed

6.2.1 UniProt

A curated protein database. See About UniProt for more details.

6.2.2 NEBcutter

New England Biolabs cutter (NEBcutter) is an online tool for emphin silico simulated digestion of DNA sequences to provide a report of which enzymes in the REBASE database will cut the DNA sequence input.[20] The tool can produce a range of reports including restriction enzyme maps, theoretical digests and links to the identified enzymes. As of version 2.0 there is an included ORF identification and editing tool based on the simulated digestion.[20] Options for restriction enzymes used include all known enzymes, subsets which are commercially available as well as sets which produce compatible termini.[20] Such information can be applied in numerous fields ranging from recombinant protein production, experimental enzyme selection, as well as gene identification and function prediction. Given these applications NEBcutter constitutes a flexible tool for simulating protein digestion which a bioinformatician should consider whenever such functionality is required.
[20]

6.2.3 ORFfinder

A tool to identify open reading frames (ORFs) from FASTA files. See ORF-FINDER: a vector for high-throughput gene identification for more details.

6.3 Tool Applications

6.3.1 UniProt

See About UniProt.

6.3.2 NEBcutter

6.3.2.1 Gene Discovery in DNA Sequences

Using the ORF functionality of NEBcutter, one is able to identify, sort, annotate and edit lists of potential ORFs from a DNA sequence.[21] These genes can then be aligned to find similar structures within a set of genes with known functions, allowing hypotheses to be formed about the function of each ORF.[21] This can be particularly useful for identifying ORFs conserved across species as well as for inferring the function of newly discovered genes. For example, identifying potential ORFs in an antibiotic resistance plasmid can be used to determine the extent of horizontal gene transfer in a population.[21]

6.3.3 ORFfinder

See ORF-FINDER: a vector for high-throughput gene identification.

Chapter 7

Protein Structure Prediction

7.1 Concepts Discussed

7.1.1 Experimental 3D Structure Determination

See lecture notes.[22]

7.1.1.1 X-Ray Crystallography

See lecture notes.[22]

7.1.1.2 NMR Spectroscopy

See lecture notes.[22]

7.1.2 Background

7.1.2.1 Definitions

See lecture notes.[22]

7.1.2.2 Protein Structures

See lecture notes.[22]

7.1.2.3 Secondary Structures

See lecture notes.[22]

7.1.3 Properties of Stable Tertiary Structure

See lecture notes.[22]

7.1.3.1

7.2 Tools Discussed

7.2.1 RCSB PDB

This database was originally constructed in the 1980s to serve as an archive for biological macromolecular crystal structures.[23] It has since been updated to include structures determined via nuclear magnetic resonance imaging (NMR) and implemented as a web-based service freely accessible to both researchers in the field and

the general public.[23] From it, a wealth of information about proteins sequence, primary, secondary and tertiary structures can be gathered to facilitate comprehensive structural analysis for proteomics applications.[23]

7.2.2 FirstGlance in Jmol

Jmol is software develop in the java language for molecular modelling chemical structures in 3D. It is available as an OS application, a JavaApplet for integration into other Java based apps, and in a JavaScript only versions using HTML5 to run on computers without java.[24] FirstGlance in Jmol is a browser based implementation of the Jmol software which is specialized for quick, widely accessible 3D visualizations of proteins, DNA, RNA from their PDB identification code.[24] The FirstGlance in Jmol web-app is integrated with the RCSB-PDB to directly fetch the PDB file of a given sequence with its protein ID.[24]

7.2.3 SWISS-MODEL

SWISS-MODEL is a homology modeling tool which uses a web-server to automate the modeling work-flow for use by researchers lacking specific computational expertise.[25] It allows input of a variety of sequence file types (FASTA, Clustal, PDB, etc.) to enable homology modeling of individual protein targets, protein targets and their templates, or target-template alignments for visualization and interpretation.[25] As homology modeling is currently the most accurate method for creating 3D protein structure models, SWISS-MODEL's accessibility and ease of use allows delivery of complex computation modeling tools to whomever may required it.

7.2.4 PDBeFold

Protein Data Bank in Europe's Fold tool (PDBeFold) allows pairwise or multiple comparison and 3D alignments of protein structures.[26] Unlike sequence alignment, the algorithm compares the geometric location of amino-acid residues between two sequences and therefore residue type is irrelevant.[26]

7.2.5 PSIPRED

PSI-blast based secondary structure PREDiction (PSIPRED) protein sequence analysis workbench is an online tool used to investigate protein secondary structures.[27] It is used for *ab initio* prediction of secondary structures form primary protein sequence data by applying different machine learning based algorithms.[27] At this time the web service is able to predict alpha helices, beta sheets, and coils but will likely be improved over time to yield more complex predictions and better accuracy.[27]

7.3 Tool Usage

7.3.1 RCSB PDB

See source.[23]

7.3.2 FirstGlance in Jmol

See source.[24]

7.3.3 SWISS-MODEL

The work-flow for 3D protein modeling in SWISS-MODEL is as follows:

1. **Input data**
   Input an amino acid sequence for the target protein. SWISS-MODEL accepts such files in FASTA, Clustal, plain text or as a UniPortKB accession code.[25]

2. **Template search**
   The data input from 1 is used as a search query for evolutionarily related protein structures in the SWISS-MODEL template library, SMTL. Two database search methods are available, the first with BLAST to provide speed with reasonable accuracy; the second is HHBlitz, which increases accuracy and sensitivity for more distantly related protein structures.[25]

3. **Template selection**
   Templates returned by the search are ranked according to quality—estimated by Global Model Quality Estimate and Quaternary Structure Quality Estimate.[25] Top-ranked templates are compared to verify they are not alternative conformational states or different regions of the target protein.[25] Options for alternative templates are displayed in a table with information about each; additionally, interactive graphical views allow visual interpretation of the available templates.[25]

4. **Model building**
   The model is built by comparing co-ordinates for each selected template to conserved co-ordinates in the target-template alignment, thereby generating a 3D framework on which to build the model.[25] Insertions/deletions in the alignment are included by loop modeling, with the final structure constructed from the remainder of non-conserved alignment positions.[25]

5. **Model quality estimation**
   Model errors are estimated using the QMEAN scoring function to generate global and per residue quality estimates.[25] This is functionally similar to a 2D scoring matrix, though the math is much more complex.

7.3.4 PDBeFold

See source.[26]

7.3.5 PSIPRED

See source.[27]

7.4 Tool Applications

7.4.1 RCSB PDB

See source.[23]

7.4.2 FirstGlance in Jmol

See source.[24]

7.4.3 SWISS-MODEL

7.4.3.1 Rational Drug Design

While 3D modeling proteins has a wide range of applications, one of particular personal interest in rational drug design. Designing drugs to interact with specific protein targets would enable rapid development of new pharmaceuticals while drastically reducing the cost of research and development.[28] 3D modeling drug and target proteins is essential as a first step towards this goal, as accurate 3D models are required to predict interactions between a drug and its target. SWISS-MODEL has a role to play in this process, but many of the other tools listed here will also make a contribution to achieving this end.[28]

7.4.4 PDBeFold

See source.[26]

7.4.5 PSIPRED

See source.[27]

Chapter 8

Secondary Structure Prediction

8.1 Concepts Discussed

8.1.1 Types of Prediction Methods

See lecture notes.[29]

8.1.1.1 Statistical/Probabilistic

See lecture notes.[29]

8.1.1.2 Knowledge-Based Methods

See lecture notes.[29]

8.1.1.3 Machine Learning Methods

See lecture notes.[29]

8.1.2 Deriving Parameters

See lecture notes.[29]

8.1.2.1 Training Dataset

See lecture notes.[29]

8.1.2.2 Testing Dataset

See lecture notes.[29]

8.1.3 Assessing Accuracy of Prediction Programs

See lecture notes.[29]

8.1.3.1 Q3

See lecture notes.[29]

8.1.3.2 Mathews Correlations Coefficient

See lecture notes.[29]

8.1.3.3 Fractional Overlap Segments

See lecture notes.[29]

8.1.4 Compositional Preferences

See lecture notes.[29]

8.1.5 Chau-Fasman Algorithm

See lecture notes.[29]

8.1.6 Nearest-Neighbour Methods

See lecture notes.[29]

8.1.7 Other Secondary Structures

See lecture notes.[29]

8.1.7.1 Trans-membrane Proteins

See lecture notes.[29]

8.1.7.2 Coiled-Coil Structures

See lecture notes.[29]

8.1.7.3 Single Stranded RNA Molecules

See lecture notes.[29]

8.1.8 Types of Nucleic Acid Secondary Structure

See lecture notes.[29]

8.1.8.1 Stems

See lecture notes.[29]

8.1.8.2 Loops

See lecture notes.[29]

8.1.8.3 Pseudo-knots

See lecture notes.[29]

8.1.8.4 Double Helix

See lecture notes.[29]

8.2 Tools Discussed

8.2.1 NCBI Nucleotide Database

Access to the NCBI Nucleotide Database is available here. The service provides an integrated search query across major nucleotide databases including NCBIs own nucleotide database and GenBank divisions storing expressed sequence tags and genome sequence survey data.[30]

8.2.2 mFold

Access to the mFold tool is available here. The web service was developed by Michael Zucker, a professor of mathematics at Rensselaer Polytechnic Institute, and is used to predict secondary structures from primary sequence data utilizing a mainly thermodynamic model.[31] The algorithm applied in this tool provides an estimate of minimum free energy conformations, thus estimating the most stable configurations from local calculated local minima..[31] If the tool is correct, the selected local minima will correspond to the global minima for a given sequence.[31]

8.3 Tool Usage

8.3.1 NCBI Nucleotide Database

See documentation.[30]

8.3.2 mFold

The mFold pipeline is composed of a number of separate applications which have been integrated to predict nucleic acid folding, hybridization and melting temperatures.[31] Instruction for using the service are:

1. **Input**
   Input is supplied via the text-box and must be a nucleic acid sequence; this input is regexed to automatically remove extraneous characters and fix case issues.[31] IUPAC codes for incomplete specification of bases can be specified, see documentation for more details.[31]

2. **Constraints**
   The text area in the constraints box allow use of optional folding constraints to increase the specificity of a query.[31] Constraints include forcing specific structures or prohibiting them for a given subsequence of base-pairs; full information about the accepted codes is available in source material.[31]

3. **Folding Parameters**
   Additional folding parameters include specifying linear vs circular ssRNA or DNA, specifying the temperature at which folding is expected to occur, as well as ionic conditions.[31] Fine control like this can enable study of secondary structure under different environmental condition and may provide information about mechanism which enable or disable a given molecule.

4. **Output Parameters**
   These can be used to specify the size and format of the images output by mFold.[31] Features such as gridlines and molecule outlines can help simplify interpretation of outputs for long sequences.[31]

5. **Folding Results**
   Folding results can be displayed in browser, or more commonly, a link can be sent to the user on completion of the batch job.[31] Images of the secondary structures include base colouring to indicate the confidence in each base-paring predicted.[31] Other output values provide overviews of the thermodynamic results of the analysis as well as allowing format customization, structure dotplots and other useful bits of information.[31]

8.4 Tool Applications

8.4.1 NCBI Nucleotide Database

See documentation.[30]

8.4.2 mFold

8.4.2.1 Investigating the Secondary Structure of Single Stranded Viral Nucleic Acids

Similar to protein structure prediction, the secondary structure of an RNA sequence can be used to infer properties of high level organization.[32] Specifically, secondary determines tertiary structure which subsequently determines function of a given nucleic acid molecule.[32] Exploring the structural motifs displayed in secondary structure may provide insights into how single-stranded nucleic acids made within a host cell influence the function of the cell, enabling viral high-jacking of the cellular machinery to reproduce. Understanding the function of such macromolecules can provide insight into the pathways utilized by a given viral species as well as providing potential targets for intervention in viral infections.

Chapter 9

Gene Expression Analysis

9.1 Concepts Discussed

9.1.1 Detecting Gene Expression

See lecture notes.[33]

9.1.2 Expression Analysis

9.1.2.1 DNA Microarrays

See lecture notes.[33]

9.1.2.2 Serial Analysis of Gene Expression (SAGE)

See lecture notes.[33]

9.1.2.3 Minimal Information About a Microarray Experiment (MIAME)

See lecture notes.[33]

9.1.2.4 Cluster Analysis

See lecture notes.[33]

9.1.2.5 2D Gel Electrophoresis

See lecture notes.[33]

9.1.2.6 Protein Microarrays

See lecture notes.[33]

9.1.2.7 Mass Spectrometry (MS)

See lecture notes.[33]

9.1.3 Expression Data Preparation

See lecture notes.[33]

9.1.3.1 Background Correction

See lecture notes.[33]

9.1.3.2 Normalization

See lecture notes.[33]

9.1.3.3 Transformation

See lecture notes.[33]

9.1.3.4 Differential Expression

See lecture notes.[33]

9.2 Tools Discussed

9.2.1 NCBI Gene Expression Omnibus (GEO)

The NCBI's GEO is an international resource which stores micro-array, next-generations sequencing and other forms of high-throughput genomics data. The repository is available online where it can be accessed publicly.[34] All submissions to GEO must follow MIAME criteria as well as organize data into platform, sample and sequence records.[34] GEO curators use records to assemble DataSets which are biologically and statistically comparable; these are used to derive profiles, which collate information across samples for an individual gene.[34] Additionally, DataSets are the input for a range of analysis and visualization tools available within the GEO web-server.[34]

9.2.2 GEO2R

GEO2R is an interactive web tool which uses submitter supplied and processed data tables for comparison of two or more sample.[35] Unlike other GEO's other DataSet analysis tools, GEO does not rely on curated GEO data, instead affecting the original data file directly.[35] Using the R packages GEOquery and limma from the Bioconductor project the tool outputs the results of statistical tests from the R language to identify deferentially expressed genes.[35]

9.3 Tool Usage

9.3.1 NCBI GEO

See documentation.[34]

9.3.2 GEO2R

The operation of GEO2R requires the following steps:

1. **Enter Series accession numbers**
   Input into GEO2R is accepted in the form of series accession numbers, representing the ordered list of all GI numbers for a specific sequence.[35] These can be input individually into text boxes or uploaded as a Series record.[35]

2. **Define Sample groups**
   Once sample are uploaded, the user must define the desired sample groupings for the exppression analysis.[35] Each grouping will get its own colour and name.[35]

3. **Assign Samples to each group**
   Samples can be assigned to groups by highlighting the associated sample row and clicking the desired group name.[35] If completed correctly, the selected samples will be coloured according the same as the selected group.[35]

4. **Edit options and features**
   These are set by GEO2R as a default, but modifications can be made by advanced users desiring more control over the analysis.[35] Options include adjusted p-values, transforming data to meet the requirements of a statistical test, or selecting which annotations to display with the results.[35]

5. **Perform the test**
   Once samples have been sorted into groups, the analysis can be run by navigating back to the GEO2R tab and clicking the "Top 250" results button.[35]

6. **Interpret the results**
   After completion of the analysis a table is displayed containing the results.[35] Each sample gets a row in the table, and clicking a sample expands the row to display a plot of the relative expression of corresponding samples between groups.[35] By identifying patterns in the plots one can determine co-expression and possibly determine co-regulation of genes between the sample conditions.[35] Results can also be downloaded for further testing with the R statistical platform.

9.4 Tool Applications

9.4.1 NCBI GEO

See documentation.[34]

9.4.2 GEO2R

9.4.2.1 Analysis of Co-regulation to Place Genes in Molecular Pathways

GEO2R enables visual exploration of expression trends between samples. By varying sample conditions, it is possible to compare gene expression across loci between samples.[36] Co-regulated genes—either positively or negatively—may indicate that these genes share a molecular pathways or regulation mechanism.[36] By placing genes within the larger expression systems within cells, between tissues, or even between organisms we are able to uncover new details about known pathways, identify gene and protein functions, as well as potentially discover novel molecular pathways. As complex webs of interconnection are elucidated, it becomes more easy to place other genes and proteins within them, eventually converging on a complete understanding of genes, proteins and their regulation within a given biological system. This powerful information will undoubtedly enable innovate new biological technology and medicine.

Chapter 10

Systems Biology

10.1 Concepts Discussed

10.1.1 Biological Systems

See lecture notes.[37]

10.1.2 Constructing Networks

See lecture notes.[37]

10.1.2.1 Bottom-Up Approach

See lecture notes.[37]

10.1.2.2 Top-Down Approach

See lecture notes.[37]

10.1.2.3 Middle-Out Approach

See lecture notes.[37]

10.1.3 Network Model Features

See lecture notes.[37]

10.1.3.1 Network Interactions

See lecture notes.[37]

10.1.3.2 Network Architectures

See lecture notes.[37]

10.1.3.3 Control Circuits

See lecture notes.[37]

10.1.3.4 Robustness

See lecture notes.[37]

10.1.3.5 Modularity

See lecture notes.[37]

10.1.3.6 Redundancy

See lecture notes.[37]

10.1.3.7 Bistable Switches

See lecture notes.[37]

10.1.4 Practical Aspects of System Models

See lecture notes.[37]

10.2 Tools Discussed

10.2.1 OmicsNet

OmicsNet enables researchers to create networks to visualize relationships between genes, proteins, transcription factors and metabolites in 3D space.[35] The web-tool contains a comprehensive, built-in knowledge-base allowing selection of multiple lists of molecules pertinent to a researcher's interests.[35] Other data formats accepted include molecular lists–wherein the first column is read as the input type and the second column is read as the expression or quantitative measurement—and network files with .sif, .txt and .graphml extensions.[35] The web-app is supported across multiple browsers including Chrome, Firefox and Safari with graphics generated using the WebGL Javascript API.[35] Network generation starts from inputted molecular lists, which are then compared to the selected external interaction database to build the interaction network.[35] The results are displayed in a highly customizable Javascript application which enables exploration using different network models, annotated and colourized by the user.[35]

10.3 Tool Usage

10.3.1 OmicsNet

1. **Input Types**
   Begin by selecting the desired analysis type from the gene/proteins, transcription factors (TF), miRNAs, or graph file options on the OmicsNet main page.[35] The gene/proteins option allows selection of organism species as well as an extensive list of available database ID formats for upload to the server.[35] The transcription factors option displays a similar input window species and database ID options specific to that tool.[35] This input patten continues for the remaining options, therefore input types vary only in species for analysis and databases available. The exception was the graph file option, which instead prompted us for a file to upload or example files for small (.txt) and large (.graphml) networks.[35]

2. **Network Building**

Examining the network building page allows concurrent analysis of up to three sample types. Available interaction options inculde PPI, miRNA-gene interactions (MGI), metabolite-protein interactiong (MPI) and TF-gene interactions (TGI) each of which have interaction database options. Netwrok tools are also available to filter the input data to control network size. On submit, the network results summarize generated subnetworks, lising the number of nodes, edges and seeds for each.[35]

3. **Network Viewer**

Proceeding to the network viewer, a Javascript interface is displayed a visualization the selected inputs and interactions.[35] Input types are colourized with a legend available underneath which a node explorer, to select a subset of molecules of interest, allowed searching and deleting nodes based on the direction of your analysis.[35] Edges appear to represent the relationships between nodes and can be customized for opacity to focus on each aspect respectively.[35] Three layouts—spherical, force-directed and layered—allow different geometries to explore the networks relationships. To the right a number of explorer windows appear to allow more specific network analysis.[35]

10.4 Tool Applications

10.4.1 OmicsNet

10.4.1.1 Exploring Biological Networks

The OmicsNet tool provides a powerful tool for visual exploration of many types of biological networks and molecular interactions therein.[38] Depending on select inputs and interactions, one is able to explore complex webs of interactions such as protein-protein, protein-metabolite, TF-gene, etc. Relating all of these nodes on in a single, manipulable dashboard allows deep analysis of such relationships. Annotation and colouration of nodes of interest, and further research using the bioinformatics tools can yield results too diverse for exhaustive description. Similar to GEO2R from the previous chapter, this tool will certainly empower new and innovate analysis, leading to significant hypotheses and enlightening conclusions which will continue to expand our understand of the complex interactions underlying biological systems.

Bibliography

[1] School of Biological Sciences and Applied Chemistry. *Topic 1: Sequence Alignment*. Ed. by Monica Wong. 2018.

[2] National Institute of Biotechnology Information. "BLAST Program Selection Guide". In: *National Library of Medicine* (2009). URL: https://blast.ncbi.nlm.nih.gov/BLAST_guide.pdf.

[3] Alexander Persemlidis and John Fondon. "Having a BLAST with bioinformatics (and avoiding BLASTphemy)". In: *Genome Biology* (2001). DOI: https://doi.org/10.1186/gb-2001-2-10-reviews2002.

[4] School of Biological Sciences and Applied Chemistry. *Lab 1: Sequence Alignment*. Ed. by Monica Wong. 2018.

[5] School of Biological Sciences and Applied Chemistry. *Topic 2: Dynamic Programming*. Ed. by Monica Wong. 2018.

[6] Gurmukh Singh and Khalid Siddiqui. "Microsoft Excel Software Usage for Teaching Science and Engineering Curriculum". In: *Journal of Educational Technology Systems* 37.4 (2009). DOI: 10.2190/et.37.4.e.

[7] GCF Global. "Excel 2010 Creating Complex Formulas". In: *GCF LearnFree.org* (2018).

[8] School of Biological Sciences and Applied Chemistry. *Lab 2: Dynamic Programming*. Ed. by Monica Wong. 2018.

[9] School of Biological Sciences and Applied Chemistry. *Topic 3: Database Searching*. Ed. by Monica Wong. 2018.

[10] et al. Hamosh A. "Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders". In: *Nucleic Acids Research* 33.Database issue (2004). DOI: 10.1093/nar/gki033.

[11] School of Biological Sciences and Applied Chemistry. *Lab 3: Database Searching*. Ed. by Monica Wong. 2018.

[12] School of Biological Sciences and Applied Chemistry. *Topic 4: Multiple Sequence Alignment*. Ed. by Monica Wong. 2018.

[13] F. Sievers et al. "Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega". In: *Molecular Systems Biology* 7.1 (2014). DOI: 10.1038/msb.2011.75.

[14] K. Duggan. "Culstal Omega". In: *EMBL-EBI* (2018). URL: https://www.ebi.ac.uk/seqdb/confluence/display/THD/Clustal+Omega.

[15] School of Biological Sciences and Applied Chemistry. *Lab 4: Multiple Sequence Alignment*. Ed. by Monica Wong. 2018.

[16] School of Biological Sciences and Applied Chemistry. *Topic 5: Phylogenetics*. Ed. by Monica Wong. 2018.

[17] J. Felsensetin. "fNeighbor". In: *National Health Research Institute* (2004).

[18] School of Biological Sciences and Applied Chemistry. *Lab 5: Phylogenetics*. Ed. by Monica Wong. 2018.

[19] School of Biological Sciences and Applied Chemistry. *Topic 6: Phylogenetics*. Ed. by Monica Wong. 2018.

[20] T. Vincze. "NEBcutter: a program to cleave DNA with restriction enzymes". In: *Nucleic Acids Research* 31.13 (2003), pp. 3688–3691. DOI: 10.1093/nar/gkg526.

[21] School of Biological Sciences and Applied Chemistry. *Lab 6: Gene Prediction*. Ed. by Monica Wong. 2018.

[22] School of Biological Sciences and Applied Chemistry. *Topic 7: Protein Structure Prediction*. Ed. by Monica Wong. 2018.

[23] H. M. Berman. "The Protein Data Bank". In: *Nucleic Acids Research* 28.1 (2000), pp. 235–242. DOI: 10.1093/nar/28.1.235.

[24] Eric Martz. "What is FirstGlance in Jmol". In: *Bioinformatics.org* (2016). URL: https://www.bioinformatics.org/firstglance/fgij/whatis.htm.

[25] Andrew Waterhouse et al. "SWISS-MODEL: homology modelling of protein structures and complexes". In: *Nucleic Acids Research* 46.W1 (2018), W296–W303. DOI: 10.1093/nar/gky427.

[26] Protein Data Bank in Europe. "PDBeFold Secondary Structure Matching". In: *PDBeFold* (2018). URL: http://www.ebi.ac.uk/pdbe/docs/Tutorials/workshop_tutorials/PDBefold.pdf.

[27] David T Jones. "Protein secondary structure prediction based on position-specific scoring matrices 1 1Edited by G. Von Heijne". In: *Journal of Molecular Biology* 292.2 (1999), pp. 195–202. DOI: 10.1006/jmbi.1999.3091.

[28] School of Biological Sciences and Applied Chemistry. *Lab 7: Protein Structure Prediction*. Ed. by Monica Wong. 2018.

[29] School of Biological Sciences and Applied Chemistry. *Topic 8: Secondary Structure Prediction*. Ed. by Monica Wong. 2018.

[30] Peter Cooper et al. *Entrez Sequences Quick Start*. National Center for Biotechnology Information, 2016. URL: https://www.ncbi.nlm.nih.gov/books/NBK44863/.

[31] Michael Zuker. "Mfold web server for nucleic acid folding and hybridization prediction." In: *Nucleic acids research* 31 (13 2003), pp. 3406–3415. ISSN: 1362-4962.

[32] School of Biological Sciences and Applied Chemistry. *Lab 8: Secondary Structure Prediction*. Ed. by Monica Wong. 2018.

[33] School of Biological Sciences and Applied Chemistry. *Topic 9: Gene Expression Analysis*. Ed. by Monica Wong. 2018.

[34] NCBI. "GEO Documentation". In: *Gene Expression Omnibus* (2016). URL: https://www.ncbi.nlm.nih.gov/geo/info/.

[35] NCBI. "About GEO2R". In: *Gene Expression Omnibus* (2016). URL: https://www.ncbi.nlm.nih.gov/geo/info/geo2r.html.

[36] School of Biological Sciences and Applied Chemistry. *Lab 9: Gene Expression Analysis*. Ed. by Monica Wong. 2018.

[37] School of Biological Sciences and Applied Chemistry. *Topic 10: Systems Biology*. Ed. by Monica Wong. 2018.

[38] School of Biological Sciences and Applied Chemistry. *Lab 10: Systems Biology*. Ed. by Monica Wong. 2018.