Bioinformatics Compendium

compiled by: Alexander Larsen

A rough guideline of topic refreshers and tools to help with all things bioinformatics. This compendium is far from complete and always growing. **Bolded** terms are usually tools, a list of tools, or a subcategory.

What to do when you need _____

Sequence Assembly and Alignment

- **Assembly(576)**
  - De Novo – without reference to a database, produces sometimes novel sequences.
    - Greedy algorithm assemblers
    - De Bruijin graph assembler- most popular with next gene sequencing
    - a short list of *some* De Novo assemblers
      - **https://en.wikipedia.org/wiki/De_novo_sequence_assemblers**
      - **Spades**
      - **Ray**
      - **AbySS**
      - **ALLPATHS-LG**
      - **Trinity**
    - There are some De Novo transcriptome assembly programs that are separate, for RNA-Seq. **This is the wiki list as of 2018**
      - **Blast2GO -** a annotater of the assembed transcripts -this is a complementary program
      - **SeqMan Ngen**
      - **SOAPdenovo-Trans**
      - **Velvet/Oases**
      - **Trans-AbySS**
      - **Trinity**
  - Mapping
- **Single alignment(576)**
  - Types(some)
    - Substitution matrix, chance of alignment
    - **BLOSUM45**
    - **BLOSUM50**
    - **BLOSUM62**empirically works the best**
    - **PAM**
  - BLOck sUbstitution Matrix (BLOSUM) 62
    - derived from set of aligned ungapped regions from protein familis called BLOCKS
    - calculate substition frequencies
    - positive for chemically similar substitution
    - common amino aids have low weights
    - rare amino acids have high weights
  - Assigning significance to alignment score
    - Bayesian framework
    - Classical approach

        - Extreme Value distribtuion
            - look at the probablility of a random score, if it is less likey than our alignment score then the score is considered significant. Plot all your scores vs randoms and get a distribution of these comparative scores.
            - Bayes theorem
    - Heuristic Algorithms
        - **BLAST**
            - basic local alignment search tool
            - compile a leist of high scoring words of score at least T, index database then **extend hits.**
            - A tradeoff between running time and sensitivity
            - don't extend a hit when the score falls below a specified threshold
        - **FASTA**
            - starts with exact seed matches instead of inexact matches that satisfy a threshold
            - extends like blast
            - join high sccoring seeds allowing for gaps
            - re-align high scoring matches using dynaimc programming
    - Different kinds of BLAST programs(program – Query from Database)
        - **BLASTP** – protein from protein
        - **BLASTN** – DNA from DNA
        - **BLASTX** – translated DNA from protein
        - **TBLASTN** – protein from translated DNA
        - **TBLASTX** – translated DNA from translated DNA
    - Sequence databases
        - Web portals, knowledge bases
            - **NCBI**
            - **EBI**
            - **Sanger**
        - Nucleotide sequences
            - **Genbank**
            - **EMBL-EBI** nucleotide sequence database
            - Comprise ~ 8% of the total database
        - Protein sequences
            - **UniProtKB**
- **Mutliple sequence alignment(576)**
    - Methods
        - Build phylogenetic trees
    - Algorithms
        - Progressive alignment algorithms
            - Star alignment
            - Guide tree approach- similar to phylo
        - Interative alginment algorithms
        - Dynamic programming is not feasible for larger and more reads O(n^k2^k)
    - Scoring
        - Entorypy based scores- best when we are most uncertain
        - sum of pairs – for a deterministic even, more certain(BLOSUM and PAM do this
    - Programs *incomplete list*
        - **https://www.ebi.ac.uk/Tools/msa/**
        - **Clustal omega-** guide tree based alignemnt

        - **Kalign-**large alignmetns
        - **MAFFT**
        - **MUSCLE-**fast and has good quality alignment according to 576
- **Long Sequence Alingment(776)**
    - **MUMmer System**
        - Indexing maximal unique matches to a myriad of large matches using preprocessed strings. Then extend these strings. Do normal substitution matrix scoring afterward to fill in some of the gaps
        - Suffix tree
        - Comparative models and operating time(fastest to slowest)
            - *LIS*- Longest increasing subsequence
            - Suffix tree
            - **Smither-Waterman**
            - FASTA -dead last by a couple orders of magnitude
    - **LAGAN**(slightly better at covering alignment compared to MUMmer
        - Three step method using 10-mer alignment allowing one mismatch
        - utilizes a trie to represent all the 3-mers of the sequence
- **Multiple Whole Genome Alignment(776)**
    - **MLAGAN**
        - requires phylogenetic tree
        - Greedy solution with local refinement
    - **Mercator**
        - Define probablistic model to solve globally
        - Inference is intractable, resort to approximations

Phylogenetic Trees

- Methods
    - Distance-based
        - UPGMA – often incorrect because ultrametric notion of distance overfits
        - Neighbor joining/nearest neighbor – unrooted trees
        - Assume additivity and sometims a “molecular clock”
    - Alignment-based methods
        - Parsimony – weighted
            - many more methods than graph search but
            - hill-climbing
            - Branch and bound
        - Probabilistic
            - so this seems to be what all the programs actually utilize, Bayes and maximum likelihood
            - felsensteins algorithm
    - Rooting a tree(afterward)
        - use a speciest that is distantly related enough to show the fork
- Programs
    - **PAML** – maximum likelihood
    - **BEAST2** – Bayesian
    - **phytools** – maximum likelihood
    - **COUNT –** maximum Parsimony, maximum likelihood
    - **ANGES –** Local Parsimony
    - **https://en.wikipedia.org/wiki/List_of_phylogenetics_software**

- **https://en.wikipedia.org/wiki/List_of_phylogenetic_tree_visualization_software**

Biological Networks

- Molecular networks (Omic networks)
  - Physical Networks
    - Transcirptional regulatory networks(overlap between metabolic network modeling tools)
      - Nodes – regulatory protein like a TF or target gene
      - Edges -TF A regulates C
      - Directed, signed, weighted graph
      - **BioCyc**
    - Protein- protein
      - Vertices – proteins
      - Edges – Protein U physically interacts with protein X
      - Undirected graph
    - Signaling networks
      - Vertices – enzymes and other proteins
      - Edges – Enzyme P modifies protein Q
      - Directed graph
      - **PathLinker- prediction algorithms**
      - **Literome**
      - **Chilibot**
      - **iHOP**
      - **eQTL electrical diagrams**
      - **HotNet – random walks/ network diffusion/ circuits**
    - Alternative pathway identification papers
      - Physical Network http://online.liebertpub.com/doi/abs/10.1089/1066527041410382
      - Maximum Edge Orientation http://nar.oxfordjournals.org/content/39/4/e22.full
      - Signaling ane Dynamic Regulatory Events Miner http://www.genome.org/cgi/doi/10.1101/gr.138628.112
      - Steiner forest - http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002887
      - Omics Integrator - http://dx.doi.org/10.1371/journal.pcbi.1004879
      - shortest paths+ steiner tree ANAT http://msb.embopress.org/content/5/1/248
  - Functional Networks
    - Metabolic network modeling
      - Vertices – enzymes
      - Enzyme M and N share a matabolite
      - Undirected and weighted graph
      - **PathoLogic**
      - **ERGO** in combination with libraries like **MetaCyc**
      - **PathwayTools**
      - **databases**
        - Kyoto encyclopedia of genes and genomes
        - Biocyc,EcoCyc, and MetaCyc
        - BRENDA
        - BiGG
        - metaTIGER
    - Genetic interaction networks (https://en.wikipedia.org/wiki/Gene_regulatory_network)
      - Vertices – genes

        - Edges – Genetic interaction between query (Q) and gene G
        - Undirected graph
        - **SGNSim, stochastic gene networks simulator**
        - **Gillespie algorithm**
- Bayesian Networks
    - A graph which is directed and acyclic
        - hill climbing search algorithm not as good
        - ***Sparse candidate***- for larger data sets like bioinformatics
    - A set of conditional distributions
- Module Networks
    - Type of bayesian networks but Conditional probability distribution represents a cluster of genes instead of individual nodes
    - sequential update, best to cluster by 10 modules for best results
    - Outperform many basic Bayesian networks
    - **LeMoNe – Learning Module Networks**
    - **LIRNET – Learning a Prior on Regulatory Potenetial from eQTL data**
    - how to find dense subgraphs with large numbers of connection
        - **HOTNET –** A set cover approach
        - **NETBAG – Network based analysis of genetic associations**
- Dependency networks Regression
    - GENIE3 algorithm for learning a dependncy network from expression data'
    - TIGRESS
- Mutual Information
    - **ARACNE**
- General applications of Networks
    - Differential subgraph identification
        - given gene expression from disease and normal studies
        - identify pathways that are most differentially altered between conditions
    - Module detection
        - Dense subgraph identification
            - Interpretaiton of gene sets
            - Identification of novel pathways
        - Set cover based methods
    - Network information flow
        - Sparse subgraph identification
        - Interpretation of gene sets
    - Prioritization of genes

Probability of sequence observations/ Gene Finding

- HMM
    - How likely is an HMM to have generated a given sequence
        - forward algorithm
    - what is the most likely "path" for generating a sequence of observations
        - Viterbi algorithm
    - Parameter estimation: How can we learn an HMM from a set of sequences?
        - Forward backward or Baum-Welch (an EM algorithm)
- Phylo-HMM multiple sequence conserved elements in the genome
    - emmissoin is a column of a multiple sequence alignment
    - Probability of an alignment and path

    - Phastcons: a phylo-hmm for finding conserved sequenece elements
    - **MutationTaster-free**
    - **PhastCons/PHAST compgen.cshl.edu/phast/**
- ChromHMM/ Histone code HMM epigentic markers
    - used with **ChIP-seq – FASTQC**
        - file type called FASTQ which is the standard as of 2016
        - then genomic Co-ordinates uses “bam”
        - segmentation (transformation) uses “wig”
        - last is actual analysis, statistic, visualization.
- Interpolated MM
    - **GLIMMER**
        - 8th order, inhomogenous, interpolated markov chain models
        - essentially ORF classification
- Eukaryotic gene finding
    - GENSCAN HMM
    - Pair HMMs

Clustering

- Motivation
    - Exploratory data analysis
        - visualization
        - understanding general characteristics of data
    - Generalization
        - infer something about a omic set based on how it relates to other objects
    - choose which one to use
        - sense of k then use
            - Gausian or k-means
        - control for the extent of dissimilarity
            - hierarchial
        - deterministic
            - Hierarchical
- Flat
    - K-means- hard clustering algorithm
        - **sklearn import Kmeans**
    - Model-based clustering
        - Gaussian mixture models -soft clustering algorithm
            - utilizes EM algorithm to learn GMM parameters
            - **Python module sklearn import GMM**
- Hierarchical
    - Top-down (divisive)
    - Bottom up (agglomerative)
    - **python module scikit**
    - **python module SciPy**
- how to measure transcriptomes
    - microarrays- *won’t usually need these at todays cost of RNA-seq and going forward*
        - **cDNA/Spotted arrays**
            - This is hybridized usually between a control and normal on plate
        - Oligonucleotide arrays
            - uses ssDNA spanning the entire genome

        - **Affymetrix**
        - **Nimblegen**
    - Sequencing
        - **RNA-seq**
            - few drawbacks

Motif Analysis

- Learning Sequence Motif Model Using Expectation (EM) (MEME)
    - **MEME Suite*****
        - Tons of tools for motifs
- Mutual Information motif FIRE (Promoters and terminators)
    - Tons of tools at: https://molbiol-tools.ca/Promoters.htm
- Quantitative trait loci (continuous phenotypes) Gene exp and metabolite abundance *incomplete list*
    - **https://omictools.com/qtl-mapping-category**
    - **RASQUAL**
    - **WEBQT**
    - **R/qtl**
    - **Qgene**
- GWAS studies(discrete phenotypes) IE disease status is binary

Epigenomic Analysis

- Algorithms
    - ChromHMM
    - Segway: Dynamic Bayesian network
- Database
    - RegulomeDB
- Programs
    - ***CLCbio – Qiagen *** Helpful for a variety of things in gene mapping***
- Protein Interaction Quantification(PIQ)
    - **PIQ - http://piq.csail.mit.edu/download.html**
        - Eukaryotic
        - bacterial
        - Prokaryotic
    - **ROC curve – confusion matrix statistic**
    - **HINT-performs best**
    - **Dnase2TF**
    - **Neph**
    - **Wellington**
    - **CENTIPEDE**
- Gaussian bivariate
    - **https://github.com/SheffieldML/GPy**
- Combined Annotation Dependent Depletion (CADD)
    - Example of an algorithm that integrates multiple types of evidence into a single score
        - Conservation
        - Epigenetic information
        - Protein Function scores for coding variants
    - Algorithms/programs
        - **DeepSEA**
        - **DeepLIFT**

RNA-Seq and Mass Spectrometry Identifycation

- RNA
  - RNA-Seq- Reverse-transcriptase-PCR
  - multireads can be recovered
  - RSEM
    - RNA-Seq by Expectation-Maximization- a generative probabilistic model
  - Public sources of RNA-Seq data
    - Gene Expression Omnibus (GEO): http://www.ncbi.nlm.nih.gov/geo/
    - Sequence Read Archive (SRA): http://www.ncbi.nlm.nih.gov/sra
    - ArrayExpress: https://www.ebi.ac.uk/arrayexpress/
- Mass spectrometry
  - applications
    - Targeted proteomics
    - Metabolomics
    - Lipidomics
    - Quantify abundance or state of all(many) proteins
  - SEQUEST/PSM(peptides spectrum match)
    - peptide matching

RNA structure Analysis

- Hints
  - Remember that RNA higher order structures often have similar structure
  - Dynamic programming breaks if *pseudoknots*
- General Algorithms
  - Nussinov
  - Energy Minimization
    - **Mfold**
    - **RNAfold**
- Grammer
  - CFG – context free grammer
  - SCFG - stochiastic
  - Algorithms- all have parallels with vitebi/forward-backward HMM algorithms
    - how likely – The Inside algorithm
    - most proxaxle parse – Cocke- Younger-Kasami (CYK) algorithm
    - what are SCFG parameters given a grammar and a set of sequences – Inside-Outside algorithm
      - **CONTRAfold**
- Software
  - **https://en.wikipedia.org/wiki/List_of_RNA_structure_prediction_software**
    - MASSIVE list with a myriad of programs
  - **CenterFOLD**
  - **CentroidHomfold**
  - **CyloFold**

Protein Structure Prediction

- Experimentally determined by expensive methods
  - x-ray crystalligraphy
  - ruclear magnetic resonance (NMR)
  - cryo-electron microscopy
- Prediction in 3D (https://en.wikipedia.org/wiki/List_of_protein_structure_prediction_software)

    - Homology modeling
        - *Protein threading*
            - modified branch and bound
        - **IntFOLD**
        - **RaptorX**
    - Fold recognition
        - **Foldit.it**
        - **IntFOLD**
        - **RaptorX**
    - Fragment assembly
        - **Rosetta**
        - **http://boinc.bakerlab.org**
        - **Evfold**
        - **QUARK**
        - **FALCON**
    - Molecular dynamics
        - **Folding@home**
        - **http://folding.stanford.edu**
        - **Abalone**
- **Secondary structure prediction**
    - **https://en.wikipedia.org/wiki/List_of_protein_secondary_structure_prediction_programs**
    - **SPIDER2**
    - **RaptorX-SS8**
    - **s2D**
- Amino Acid Review
    - amino group
    - side chain
    - carboxyl group