CAP 5510/CGS 5166: Bioinformatics & Bioinformatic Tools

GIRI NARASIMHAN, SCIS, FIU

Course Preliminaries

- Course Webpage: **http: //www.cs.fiu.edu/~giri/teach/BioinfF18.html**
  - Lecture Slides; Reading Material; Announcements; Homework
  - VISIT OFTEN!
- Class meets 5:00 – 6:15 PM, ECS 138, MW
- Office ECS 254B; Office Hours: By Appointment Only
- Phone: x-3748; Email: giri@cis.fiu.edu
- Final Exam: Monday, 12/3/2018, 5:00 – 7:00 PM, ECS 138
- Extra 1 credit for CGS 5166 students, if needed

Syllabus

- Fundamentals of Biology, Statistics, & Bioinformatics
- Databases; Data Integration; BioPerl & BioPython;
- Sequence Alignment, Multiple Sequence Alignment Sequencing; Next Generation Sequencing & Applications
- Pattern Discovery, Learning, Prediction & Inference; Machine Learning: NN, HMM, SOM, SVM, etc.
- Gene Regulation; Regulatory Elements; & networks Transcriptomics: Analysis of Gene Expression Data; Genomics, Proteomics, Transcriptomics; other Omics
- Gene Ontology and Pathways; Protein-protein interactions Comparative Genomics
- Phylogenetic Analysis
- Structural Bioinformatics: RNA and Proteins
- Genetics and Genome-Wide Association Schemes Single Nucleotide Polymorphisms
  Misc.: Omics; Alternative Splicing; Epigenetics;
- Cancer Bioinformatics; Microbiomes and Metagenomics;
- Software Engineering; Visualization;

Evaluation

- Semester Project (45 %)
- Homework Assignments (20 %)
- Exam (15 %)
- Quizzes (10 %)
- Summary Reports of Interest (5 %)
- Class Participation (5 %)

http: //www.cs.fiu.edu/~giri/teach/BioinfF18.html

Some History …

- What major world event took place on **26 June, 2000**?
- Other dates in Bioinformatics history:
  - 1758 – work of **Carl Linnaeus** – **taxonomy**
  - mid 1800s – **Gregor Mendel** – **genetics**
  - mid 1800s – **Charles Darwin** – **evolution**
  - 1953 – Watson, Crick, Franklin **Structure of DNA**
- More important dates
  - 1975 – Sanger Sequencing
  - 1977 – first bacteriophage sequenced
  - 1978 – Dayhoff's Atlas of Protein Sequence and Structure 1980s – EMBL, GenBank, SWISSProt, and DDBJ
  - 1990 – HGP initiated
  - Oct, 2013 – first Bioinformatics Nobel Prize (Chem): **Karplus**, **Warshel**, & **Levitt** models for complex chemical processes

Algorithms & Hardware

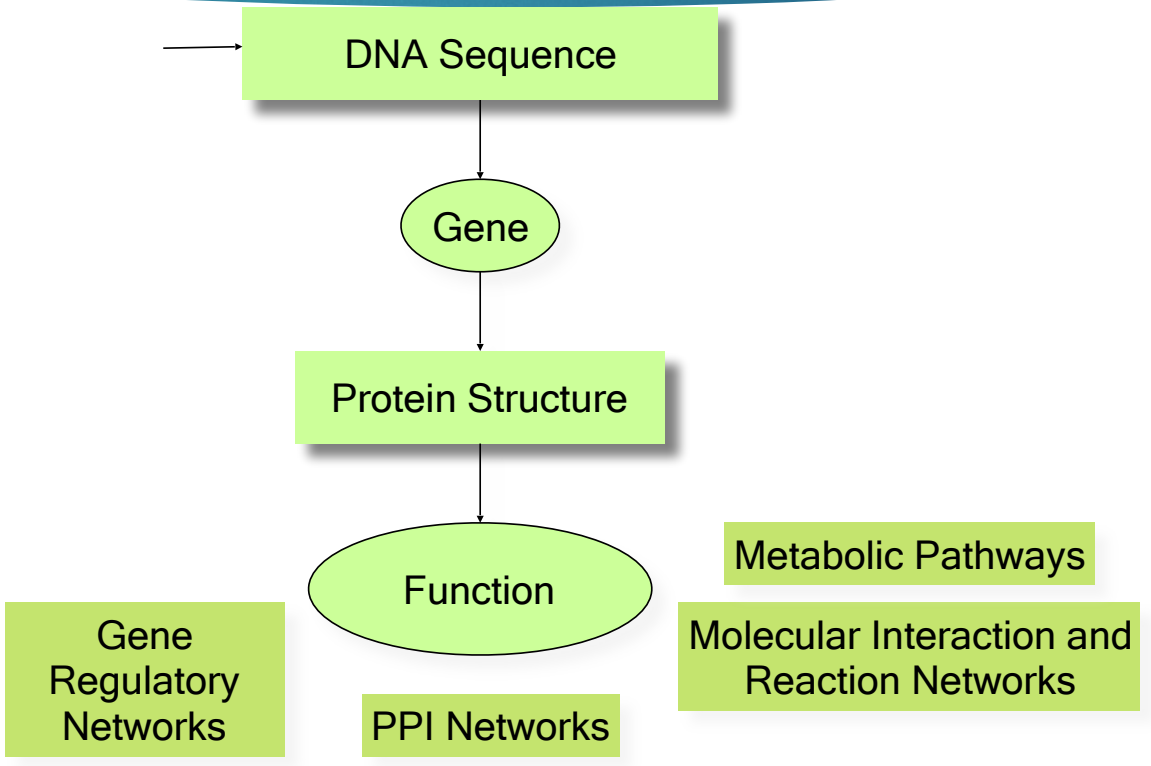
- Moore's Law Faster processors, larger and faster memory, larger external memories
- Optimization "Linear Programming is tractable"
- Convex Programming Interior Point Methods
- Energy Minimization Soft Computing Methods (Simulated Annealing, Neural Networks, ...)
- Parallel/Grid/Cloud Computing CHARMM ported to parallel environments
- GPU Computing NVIDIA video cards do more than just graphics, and can be programmed (in C/C++) to deliver on high performance scientific computing
- Quantum Computing Showed that some problems can be solved more efficiently on a quantum computer

Introduction

1. **What is Bioinformatics?**
   - Analysis of biological data with computing & statistical tools.
2. **The different aspects of Informatics?**
   - Data Management (Database Technology, Internet Programming)
   - Data Analysis (Data Mining, Modeling, Statistics)
   - Development of Efficient Algorithms
   - Visualization and Interface Design (HCI, Graphics)

1. **How to assist biological research?**
   - Build databases for data
   - Build efficient tools for search, retrieval, analysis, & visualization
   - Propose models and efficient tools to verify the model using known data
   - use predicted information to narrow down search
   - propose new experiments based on model or analysis
   - Build smart, hyperlinked, integrated mining environments

Overall Goals



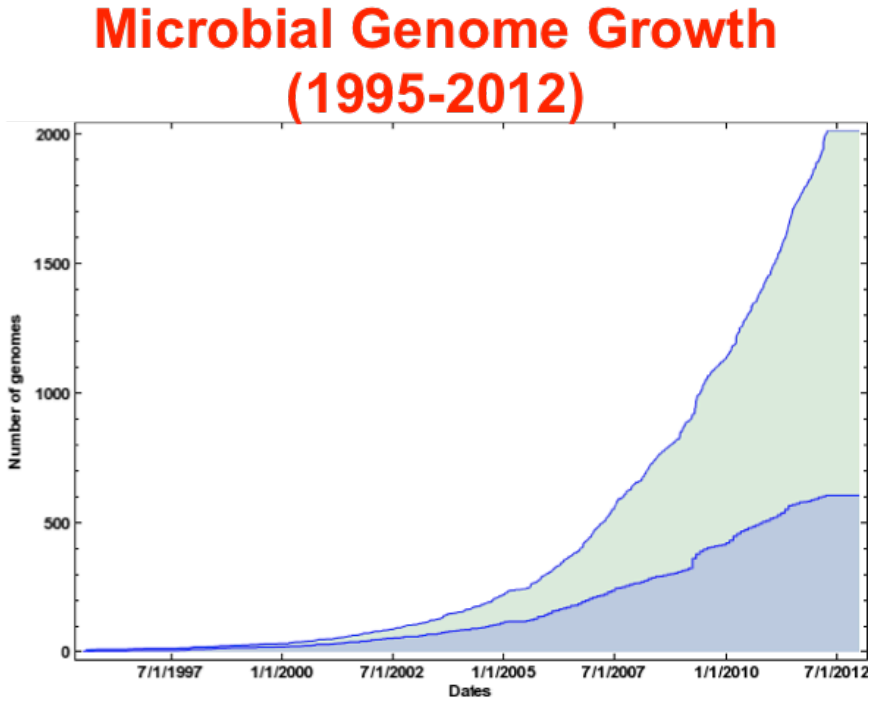
Perspectives of Bioinformatics

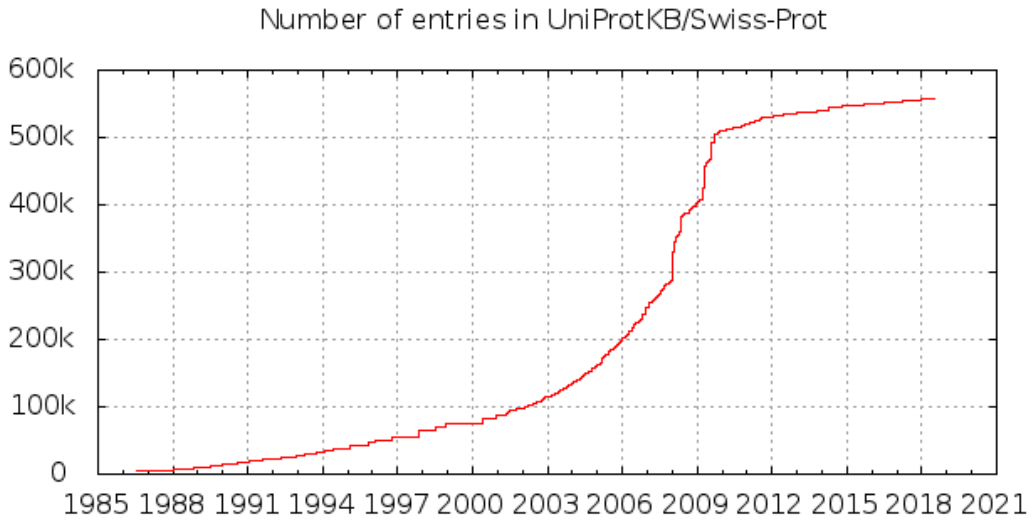
- **Molecular**: DNA, RNA, proteins, ligands, toxins, . . .
- **Cellular** chromosome, nucleus, cell wall, chromatin, organelles, organization of a single cell
- **Tissue & Organ**: Collection of cells: gene expression
- **Organism or Systems Biology**: Genome, variations within organism, or over physiological or pathological states, epigenome
- **Community**: Metagenome, Microbiome, Ecology, …
- **All life**: Tree of life, phylogeny, variations, comparative studies

Phenomenal Growth of Information

- **Life on Earth** ~8.5M eukaryotic species [Mora, C., et al., (2011). PLoS Biol, 9(8)];
- **Human Genome** has 3 billion bp with 32,000+ genes.
- **GenBank** Release 157/175/193/**205** (Dec 2006/09/12/14) contains over 64/112/161/**179** million sequence entries totaling over 69/110/126/**184** Gb from over 2,500/?/9000/**11000** organisms (Storage: 600 GBytes uncompressed); More at http://www.ncbi.nlm.nih.gov/genbank/statistics
- 435/624/3880/**30,000** complete microbial genomes sequenced of which 4500 are virus genomes.
- **UniProtKB/Swiss-Prot** Release 54.7/2012 11/**2015 01** (Jan08/Nov'12/**Jan'15**): 333K/530K/**550K** entries; 120/191/**194** million amino acids.

Phenomenal Growth of Information…2





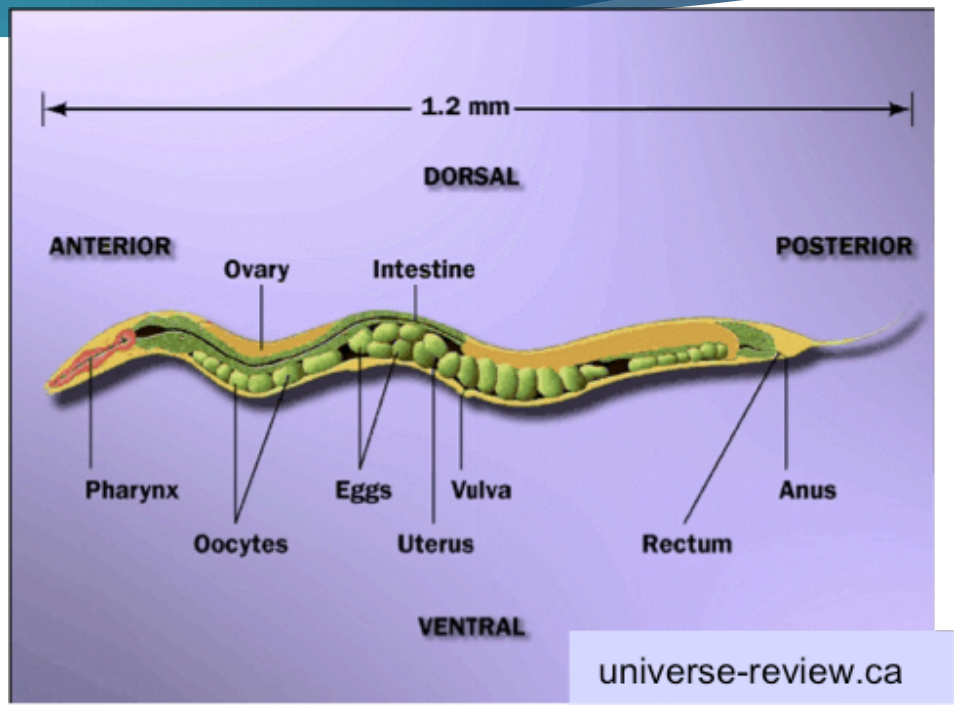
1800 Complete genomes

- Caenorhabditis elegans
- Arabidopsis thaliana
- Saccharomyces cerevisiae
- Mus musculus
- Homo sapiens
- Oryza sativa
- Plasmodium falciparum
- Drosophila melanogaster
- Anopheles gambiae
- Macaca mulatta
- Bos taurus
- Felis catus
- Gallus gallus

And Genome Sizes ...

Organism	Size	Date	No. of Genes (est.)
HIV Type I	9.2 Kb	1997	9
M. genitalium	0.58 Mb	1998	525
H. influenzae	1.8 Mb	1995	1,740
E. coli	4.7 Mb	1997	4,000
S. cerevisiae	12.1 Mb	1996	6,034
C. elegans	97 Mb	1998	19,099
A. thaliana	100 Mb	2000	25,000
D. melanogaster	180 Mb	2000	13,061
M. musculus	3 Gb	2002	30,000
H. sapiens	3 Gb	2001	32,000

Caenorhabditis Elegans

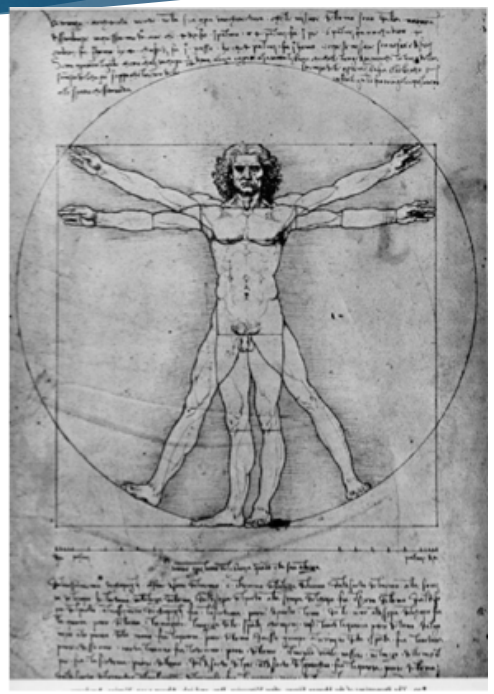


C. elegans: The Model worm

- Entire genome – 1998; 8 year effort
- 1st animal; 2nd eukaryote (after yeast)
- Nematode (phylum)
- Easy to experiment with; Easily observable
- 97 million bases; 20,000 genes;
- 12,000 with known function; 6 Chromosomes;
- GC content 36%
- 959 cells; 302-cell nervous system
- 36% of proteins common with human
- 15 Kb mitochondrial genome
- Results in ACeDB
- 25% of genes in operons
- Important for HGP: technology, software, scale/efficiency
- 182 genes with alternative splice variants

Homo sapiens

- Sequenced – 2001; 15 year effort
- 3 billion bases, 500 gaps
- Variable density of Genes, SNPs, CpG islands
- ~ 1.1% of genome codes for proteins; 99%?
- ~ 40-48% of genome consists of repeat sequences
- ~ 10 % of the genome consists of repeats called ALUs
- ~ 5 % of the genome consists of long repeats (>1 Kb)
- 223 genes common with bacteria that are missing from worm, fly or yeast.



Sequence Alignments: Why we need them?

>gi|12643549|sp|O18381|PAX6_DROME Paired box protein Pax-6 (Eyeless protein)
MRNLPCLGTAGGSGLGGIAGKPSPTMEAVEASTASHRHSTSSYFATTYYHLTDDECHSGVNQLGGVFVGG
RPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETGSIRPRAIGGSKPRVATAEVVSKIS
QYKRECPSIFAWEIRDRLLQENVCTNDNIPSVSSINRVLRNLAAQKEQQSTGSGSSSTSAGNSISAKVSV
SIGGNVSNVASGSRGTLSSSTDLMQTATPLNSSESGGASNSGEGSEQEAIYEKLRLLNTQHAAGPGPLEP
ARAAPLVGQSPNHLGTRSSHPQLVHGNHQALQQHQQQSWPPRHYSGSWYPTSLSEIPISSAPNIASVTAY
ASGPSLAHSLSPPNDIESLASIGHQRNCPVATEDIHLKKELDGHQSDETGSGEGENSNGGASNIGNTEDD
QARLILKRKLQRNRTSFTNDQIDSLEKEFERTHYPDVFARERLAGKIGLPEARIQVWFSNRRAKWRREEK
LRNQRRTPNSTGASATSSSTSATASLTDSPNSLSACSSLLSGSAGGPSVSTINGLSSPSTLSTNVNAPTL
GAGIDSSESPTPIPHIRPSCTSDNDNGRQSEDCRRVCSPCPLGVGGHQNTHHIQSNGHAQGHALVPAISP
RLNFNSGSFGAMYSNMHHTALSMSDSYGAVTPIPSFNHSAVGPLAPPSPIPQQGDLTPSSLYPCHMTLRP
PPMAPAHHHIVPGDGGRPAGVGLGSGQSANLGASCSGSGYEVLSAYALPPPPMASSSAADSSFSAASSAS
ANVTPHHTIAQESCPSPCSSASHFGVAHSSGFSSDPISPAVSSYAHMSYNYASSANTMTPSSASGTSAHV
APGKQQFFASCFYSPWV

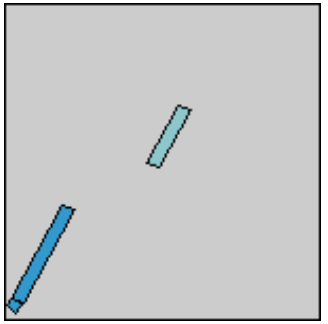
>gi|6174889|PAX6_HUMAN Paired box protein (Oculorhombin) (Aniridia, type II prote
MQNSHSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETGSIRPRA
IGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDRLLSEGVCTNDNIPSVSSINRVLRNLASEKQQMGAD
GMYDKLRMLNGQTGSWGTRPGWYPGTSVPGQPTQDGCQQQEGGGENTNSISSNGEDSDEAQMRLQLKRKL
QRNRTSFTQEQIEALEKEFERTHYPDVFARERLAAKIDLPEARIQVWFSNRRAKWRREEKLRNQRRQASN
TPSHIPISSSFSTSVYQPIPQPTTPVSSFTSGSMLGRTDTALTNTYSALPPMPSFTMANNLPMQPPVPSQ
TSSYSCMLPTSPSVNGRSYDTYTPPHMQTHMNSQPMGTSGTTSTGLISPGVSVPVQVPGSEPDMSQYWPR
LQ

Drosophila Eyeless vs. Human Aniridia

```
Query: 57   HSGVNQLGGVFVGGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETG 116
            HSGVNQLGGVFV GRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETG
Sbjct: 5    HSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETG 64

Query: 117 SIRPRAIGGSKPRVATAEVVSKISQYKRECPSIFAWEIRDRLLQENVCTNDNIPSVSSIN 176
            SIRPRAIGGSKPRVAT EVVSKI+QYKRECPSIFAWEIRDRLL E VCTNDNIPSVSSIN
Sbjct: 65   SIRPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDRLLSEGVCTNDNIPSVSSIN 124

Query: 177 RVLRNLAAQKEQ 188
            RVLRNLA++K+Q
Sbjct: 125 RVLRNLASEKQQ 136
```

```
Query: 417 TEDDQARLILKRKLQRNRTSFTNDQIDSLEKEFERTHYPDVFARERLAGKIGLPEARIQV 476
           +++ Q RL LKRKLQRNRTSFT +QI++LEKEFERTHYPDVFARERLA KI LPEARIQV
Sbjct: 197 SDEAQMRLQLKRKLQRNRTSFTQEQIEALEKEFERTHYPDVFARERLAAKIDLPEARIQV 256

Query: 477 WFSNRRAKWRREEKLRNQRR 496
           WFSNRRAKWRREEKLRNQRR
Sbjct: 257 WFSNRRAKWRREEKLRNQRR 276
```

E-Value = 2e-31



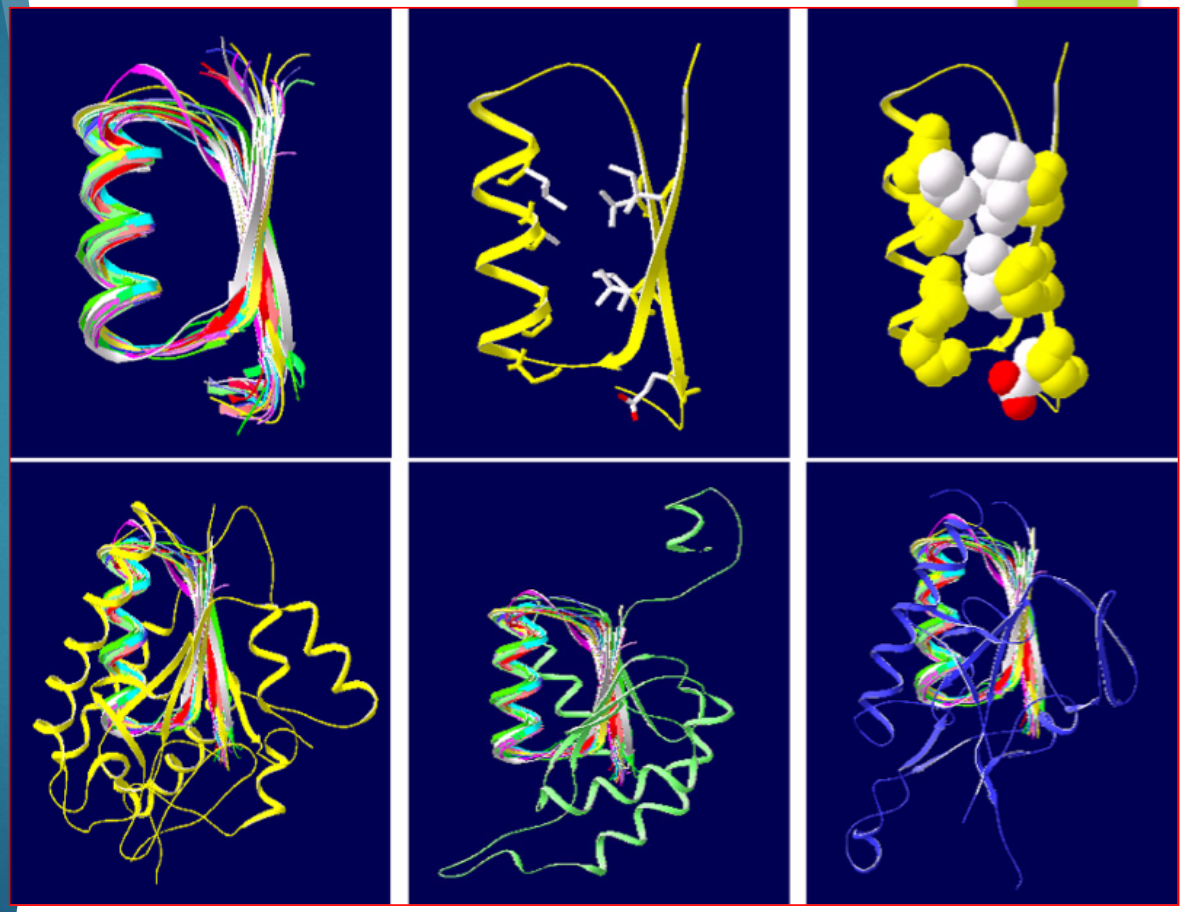
Motif Detection in Protein Sequences

- MTDKMQSLALAPVGNLDSYIRAANAWPMLSADEERALAEKLHYHGDLEAA
  KTLILSHLRFVVHIARNYAGYGLPQADLIQEGNIGLMKAVRRFNPEVGVR
  LVSFAVHWIKAEIHEYVLRNWRIVKVATTKAQRKLFFNLRKTKQRLGWFN
  QDEVEMVARELGVTSKDVREMESRMAAQDMTFDLSSDDDSDSQPMAPVLY
  LQDKSSNFADGIEDDNWEEQAANRLTDAMQGLDERSQDIIRARWLDEDNK
  STLQELADRYGVSAERVRQLEKNAMKKLRAAIEA
- MTDKMQSLALAPVGNLDSYIRAANAWPMLSADEERALAEKLHYHGDLEAA
  KTLILSHLRFVVHIARNYAGYGLPQADLIQEGNIGLMKAVRRFNPEVGVR
  LVSFAVHWIKAEIHEYVLRNWRIVKVATTKAQRKLFFNLRKTKQRLGWFN
  QDEVEMVARELGVTSKDVREMESRMAAQDMTFDLSSDDDSDSQPMAPVLY
  LQDKSSNFADGIEDDNWEEQAANRLTDAMQGLDERSQDIIRARWLDEDNK
  STLQELADRYGVSAERVRQLEKNAMKKLRAAIEA

[G. Narasimhan, et al., "Mining Protein Sequences for Motifs," *J. of Comput Biol*, 9(5):707-720, 2002.]

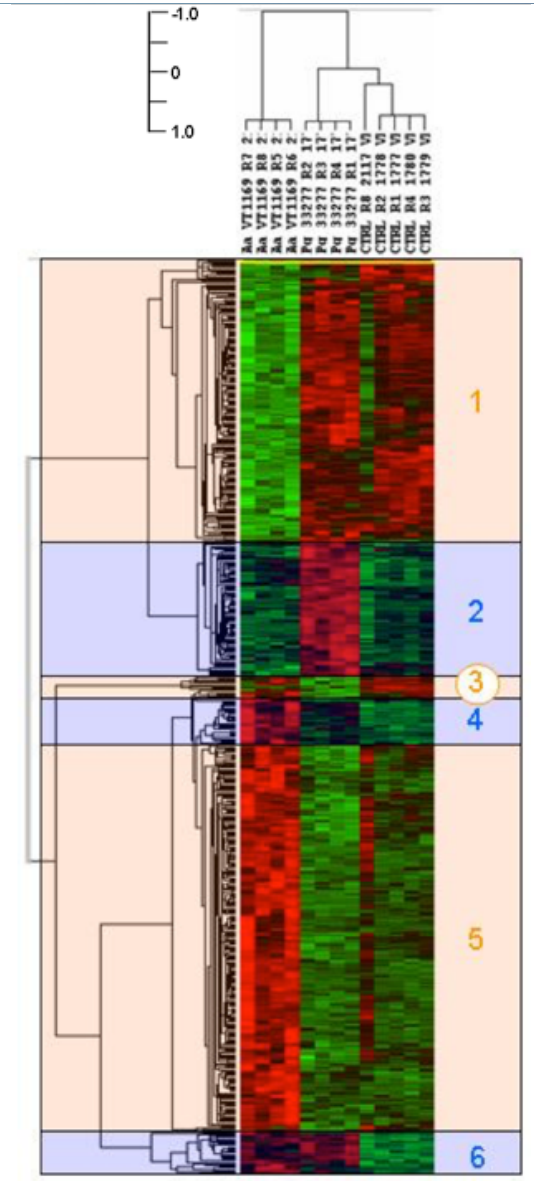
Patterns in Protein Structures

T. Milledge et al., "Sequence Structure Patterns: Discovery and Applications", CBG 2005



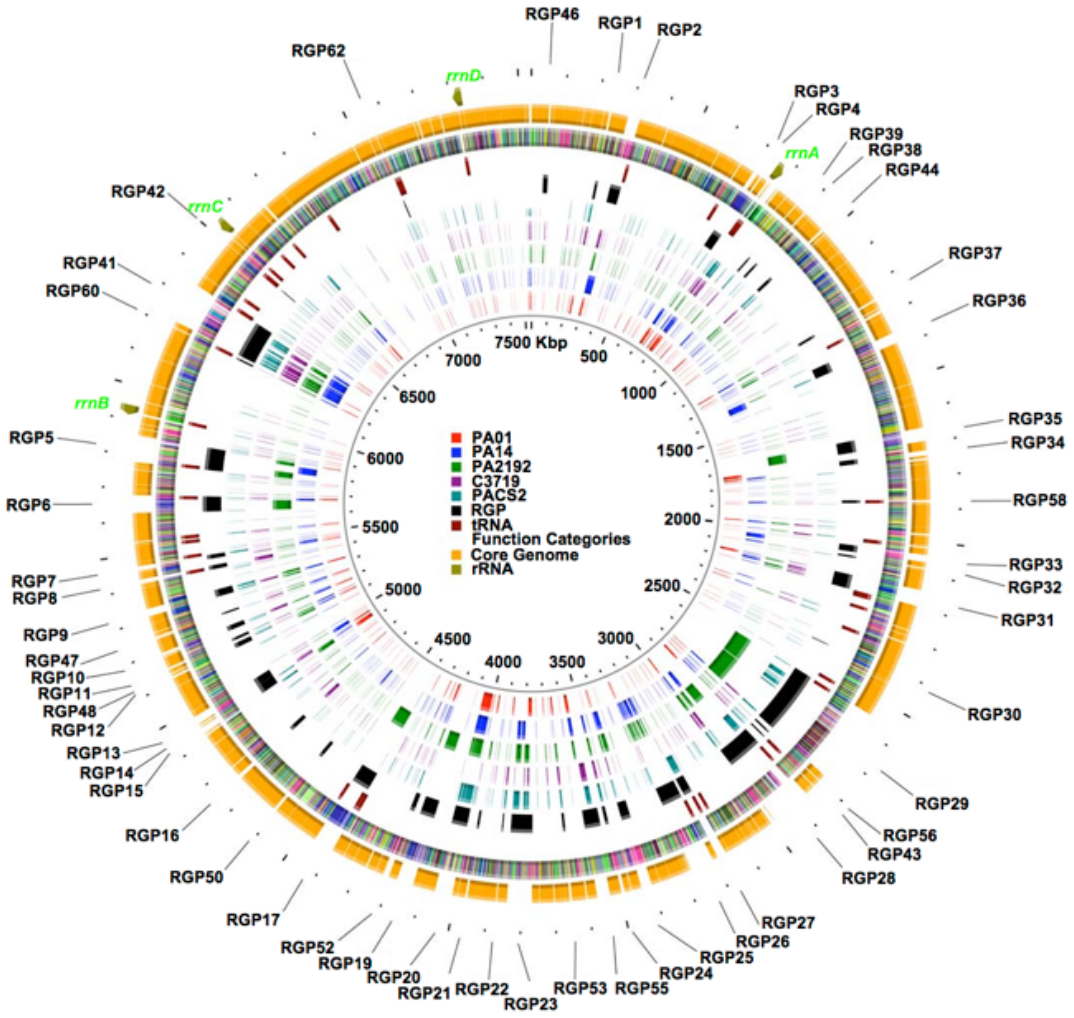
Microarray Analysis

Handfield et al., Distinct Expression Profiles Characterize Oral Epithelium-Microbiota Interaction", Cellular Microbiology, 2005



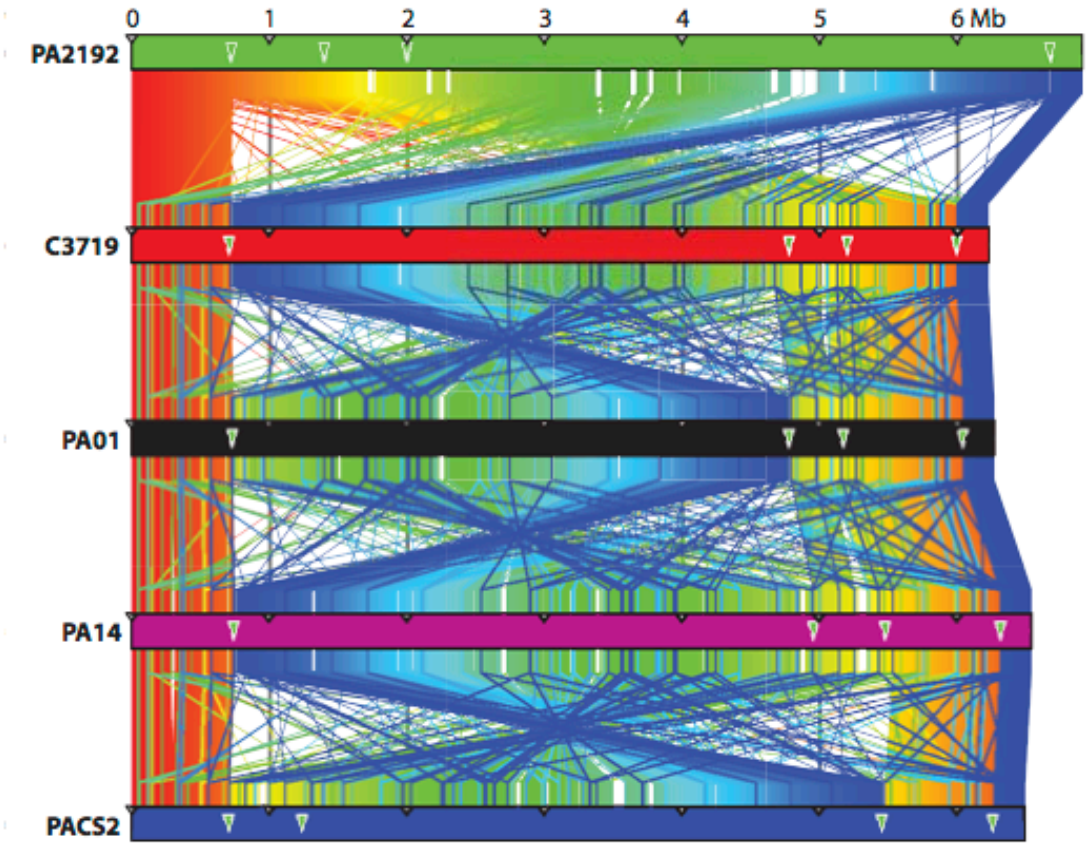
Comparative Genomics

K. Mathee, et al., "Dynamics of *Pseudomonas aeruginosa* genome evolution," Proc Natl Acad of Sciences (PNAS), 2008



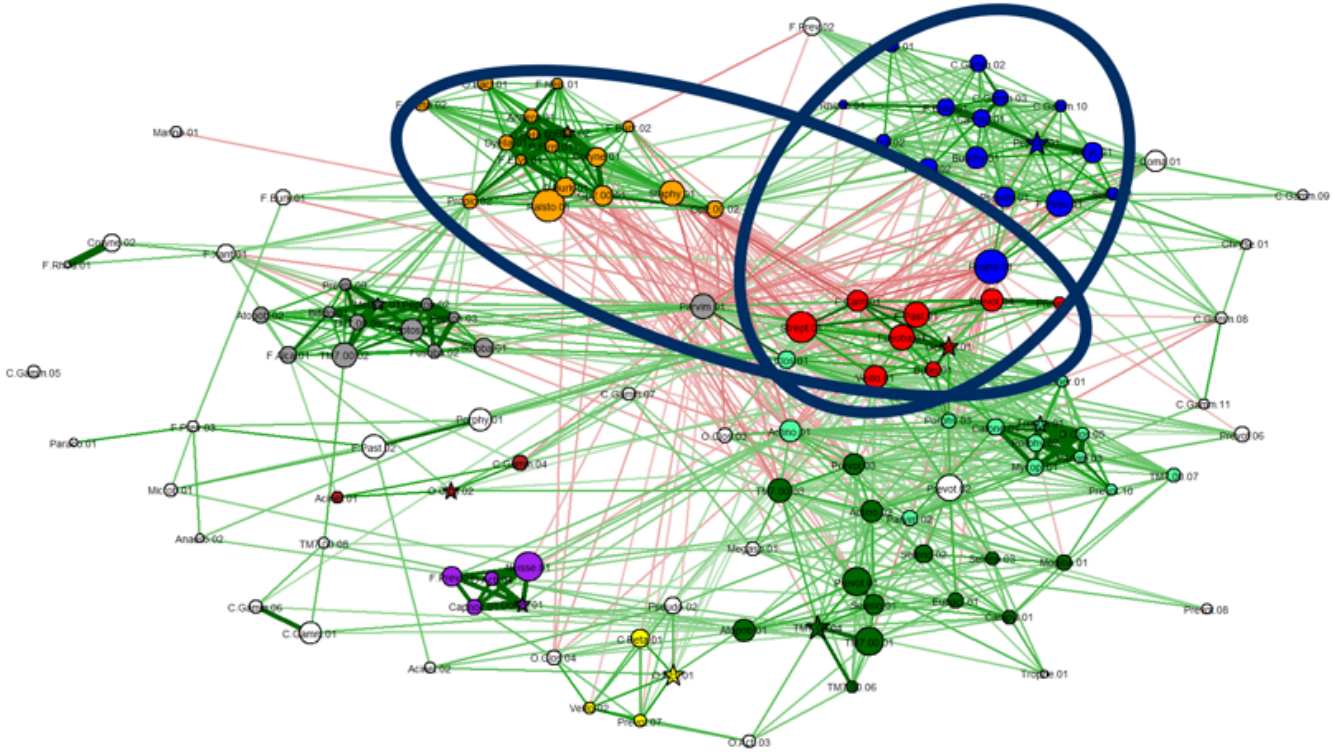
Comparative Genomics

K. Mathee, et al., "Dynamics of *Pseudomonas aeruginosa* genome evolution," Proc Natl Acad of Sciences (PNAS), 2008



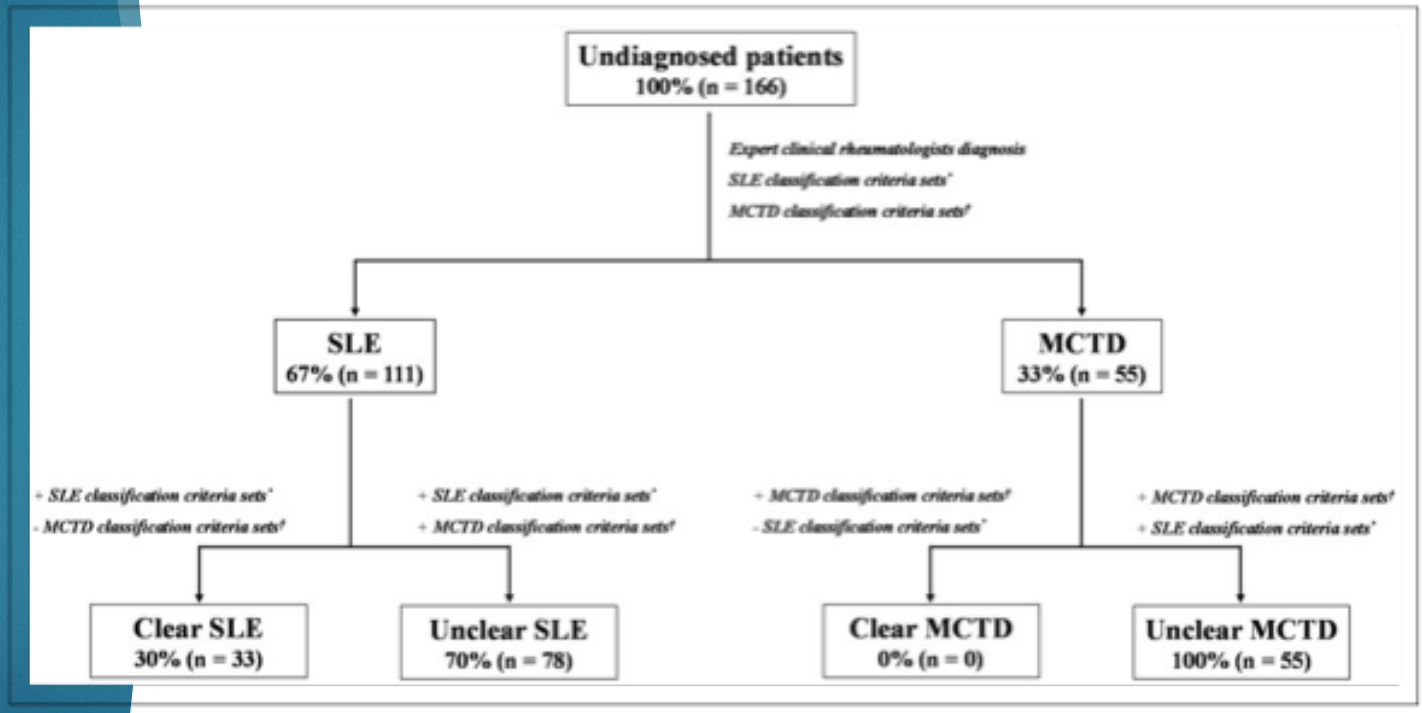
Microbiomes: The Ultimate "Social Network"

Fernandez, Riveros, Campos, Mathee, **Narasimhan Microbial "Social" Networks,** *BMC Genomics*, 2015.



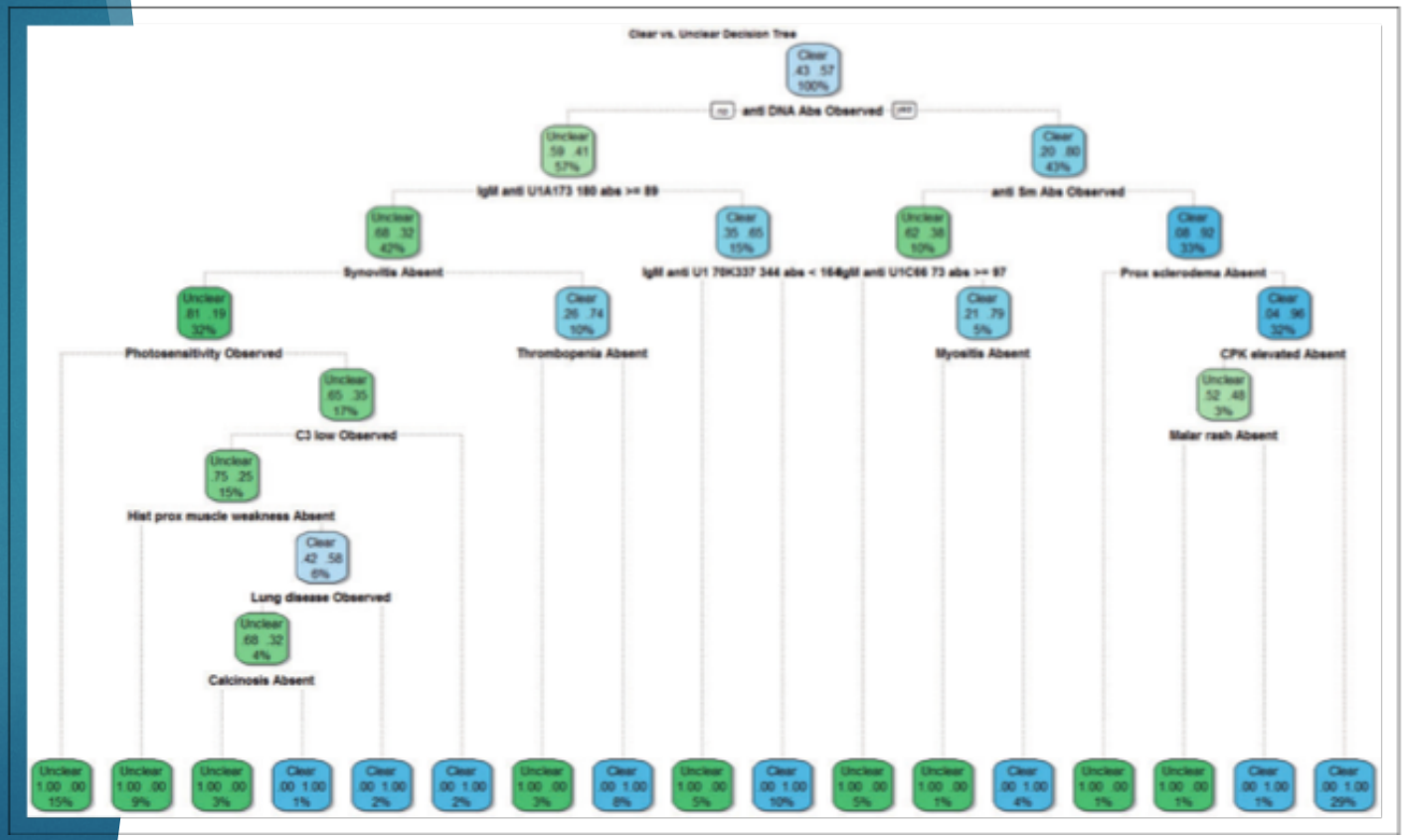
Distinguishing SLE from MCTD

State of the Art



Distinguishing SLE from MCTD with ML

Mesa, A., Fernandez, M., Wu, W., Narasimhan, G., Greidinger, E.L. and Mills, D.K., 2017. Can SLE classification rules be effectively applied to diagnose unclear SLE cases?. *Lupus*, 26(2), pp. 150-162.



Distinguishing SLE from MCTD with ML

Mesa, A., Fernandez, M., Wu, W., Narasimhan, G., Greidinger, E.L. and Mills, D.K., 2017. Can SLE classification rules be effectively applied to diagnose unclear SLE cases?. *Lupus*, 26(2), pp. 150-162.

Table 4 Evaluating novel proposed classification rule for unclear SLE and MCTD patients

Classification criteria sets		*Accuracy*	*Sensitivity*	*Specificity*
New proposed "Lu-vs-M" rule		*96.30%*	*61.54%*	*50.00%*
SLE	***SLICC***	62.96%	93.75%	18.18%
	ACR	55.55%	75.00%	27.27%
MCTD	***Alarcón-Segovia***	22.22%	18.75%	27.27%
	Sharp	48.14%	62.50%	27.27%
	Kasukawa	50.00%	80.00%	9.09%
	Kahn	29.63%	6.25%	63.64%

Analyses were performed in SPSS (version 18) and included unclear SLE ($n = 16$) and MCTD ($n = 11$) patients from the validation group. Lu-vs-M: SLE vs MCTD; SLICC: Systemic Lupus International Collaborating Clinics; ACR: American College of Rheumatology; SLE: systemic lupus erythematosus; MCTD: mixed connective tissue disease.

The SIDS Mystery

- 18000 Amish people in Pennsylvania
- Mostly intermarried due to religious doctrine
- rare recessive diseases occurred with high frequencies.
- SIDS: 3000 deaths/year (US); 21 deaths (Amish community)
- Many research centers failed to identify cause
- Collaboration between Affymetrix, TGEN & Clinic for special children solved the problem in 2 months
- Studied 10000 SNPs using microarray technology
- Their experiments showed that all the sick infants had two mutant copies of a specific gene, and their parents were carriers of the mutant gene.
- Conclusion: Disease caused by 2 abnormal copies of TSPYL gene
- Identified genes expressed in key organs (brainstem,testes)
- http://www.affymetrix.com/community/wayahead/modern_miracle.affx

The Alzheimer's Mystery

- Search for the "Alzheimer's Laboratory", an episode of 60 minutes that was aired by CBS in Nov 2016 and then again in Jan 2018.
- This is now in Homework 1. More later …