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

Optional Software App: Momentos

- Leonardo Marmol will tell us a bit about Momentos
- Survey for App use

Momentos Survey

- Survey Consent

https://users.cs.fiu.edu/~giri/Momentos/MomemtosConsentForm.pdf

- Survey link

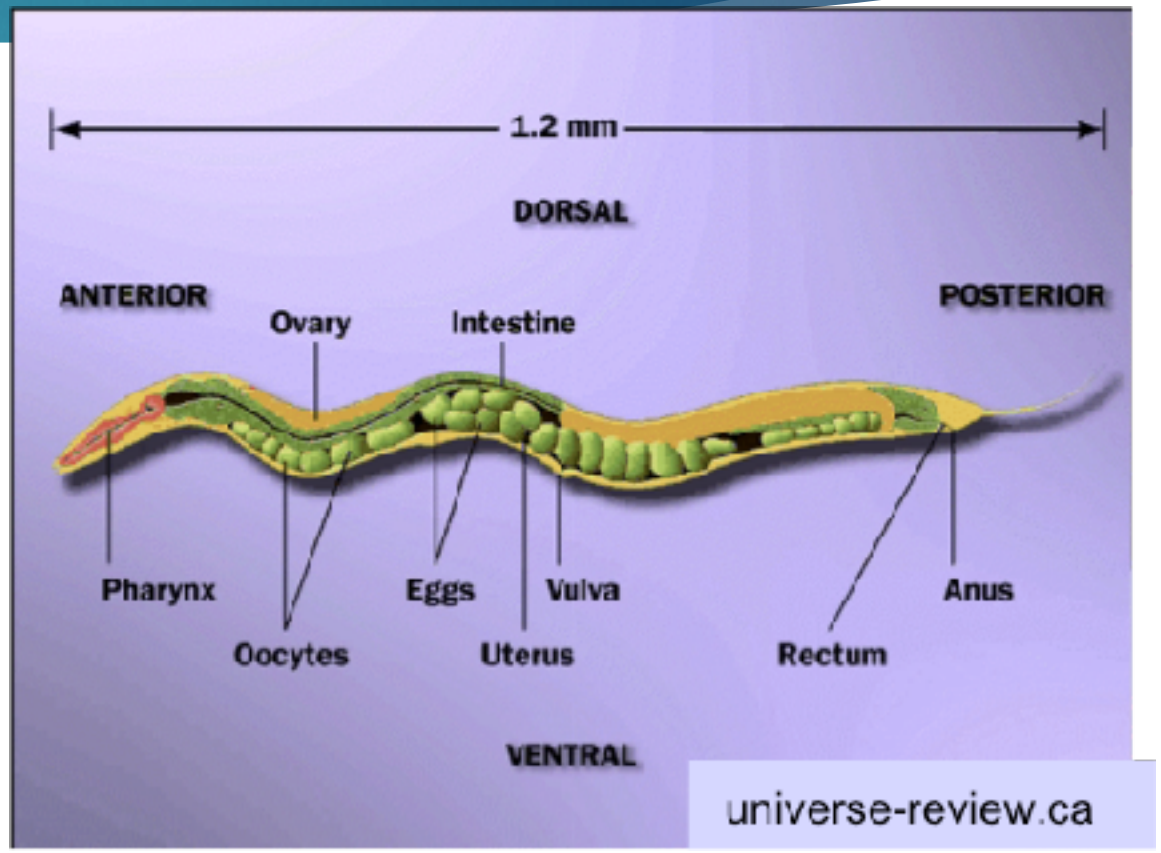
https://drive.google.com/open?id=1VG5EdJGYXRlvfzhYgIqPN8domKz6oHQHm2sQsHyrcbg

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

Caenorhabditis Elegans

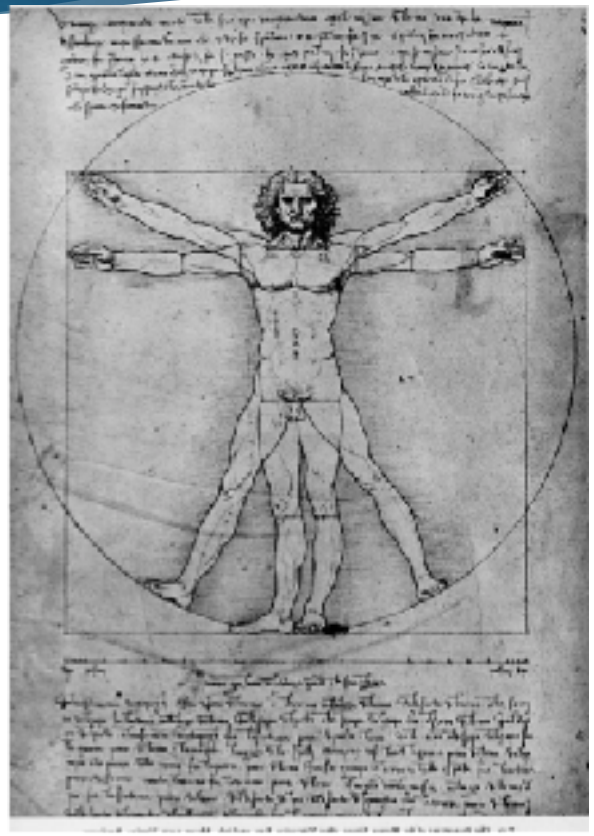


C. elegans: The Model worm

- 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
```

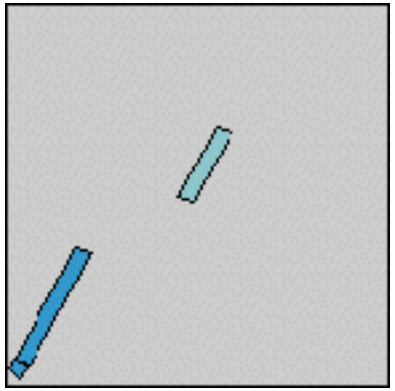
```
>gi|6174889|PAX6_HUMAN Paired box protein (Oculorhombin) (Aniridia, type II prote
MQNSHSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETGSIRPRA
IGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDRLLSEGVCTNDNIPSVSSINRVLRNLASEKQQMGAD
GMYDKLRMLNGQTGSWGTRPGWYPGTSVPGQPTQDGCQQQEGGGENTNSISSNGEDSDEAQMRLQLKRKL
QRNRTSFTQEQIEALEKEFERTHYPDVFARERLAAKIDLPEARIQVWFSNRRAKWRREEKLRNQR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



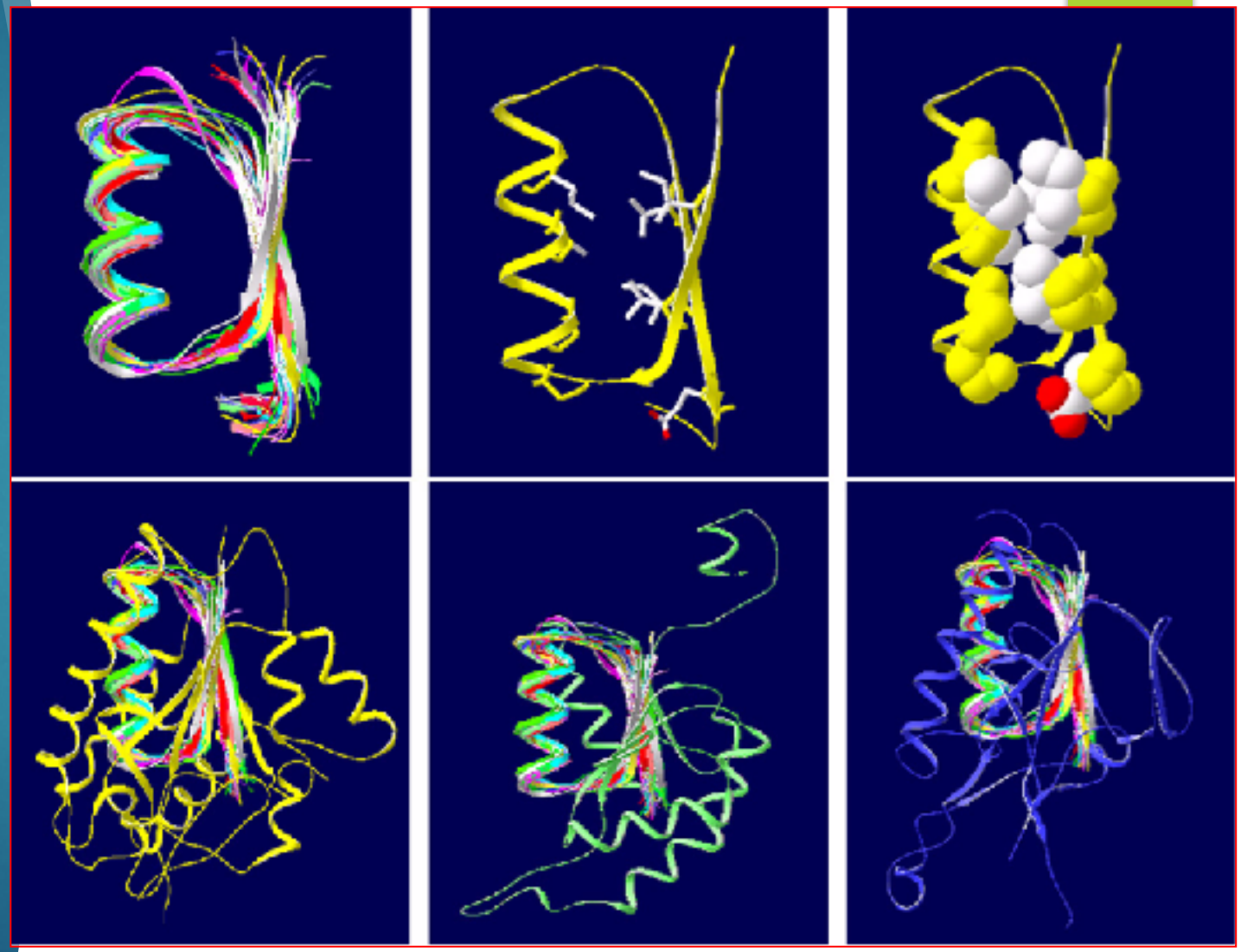
Motif Detection in Protein Sequences

- 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.]

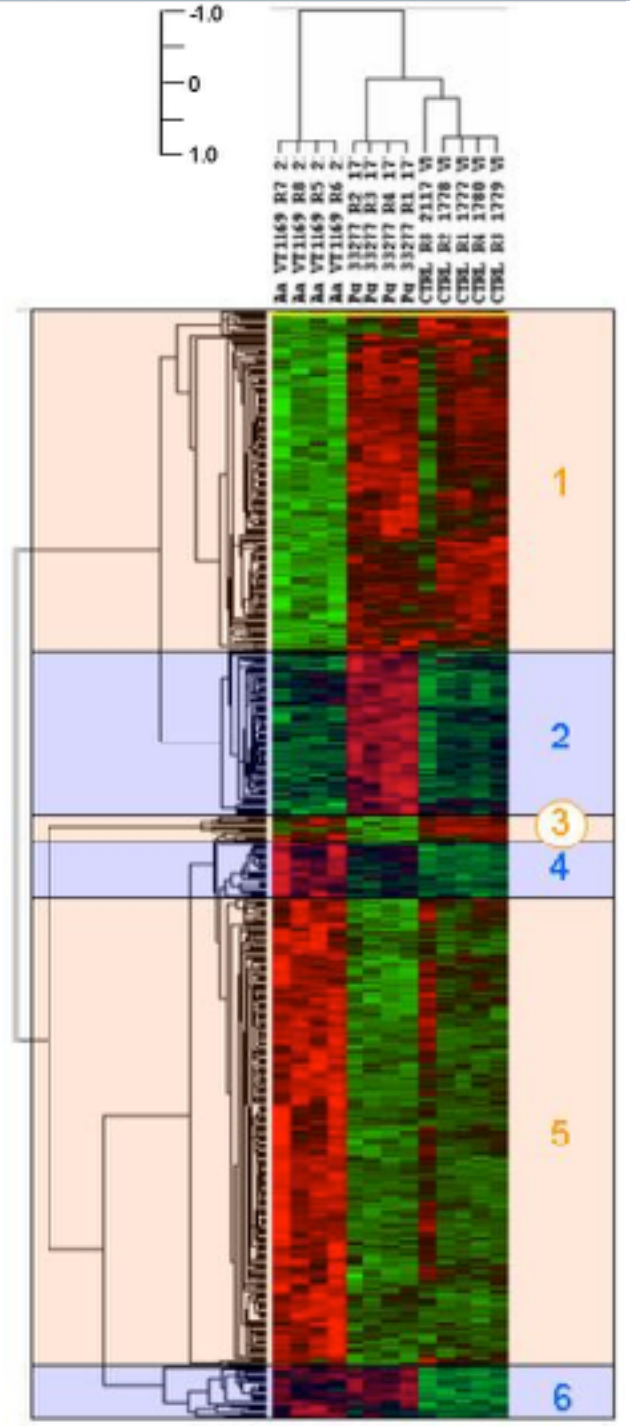
Patterns in Protein Structures

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



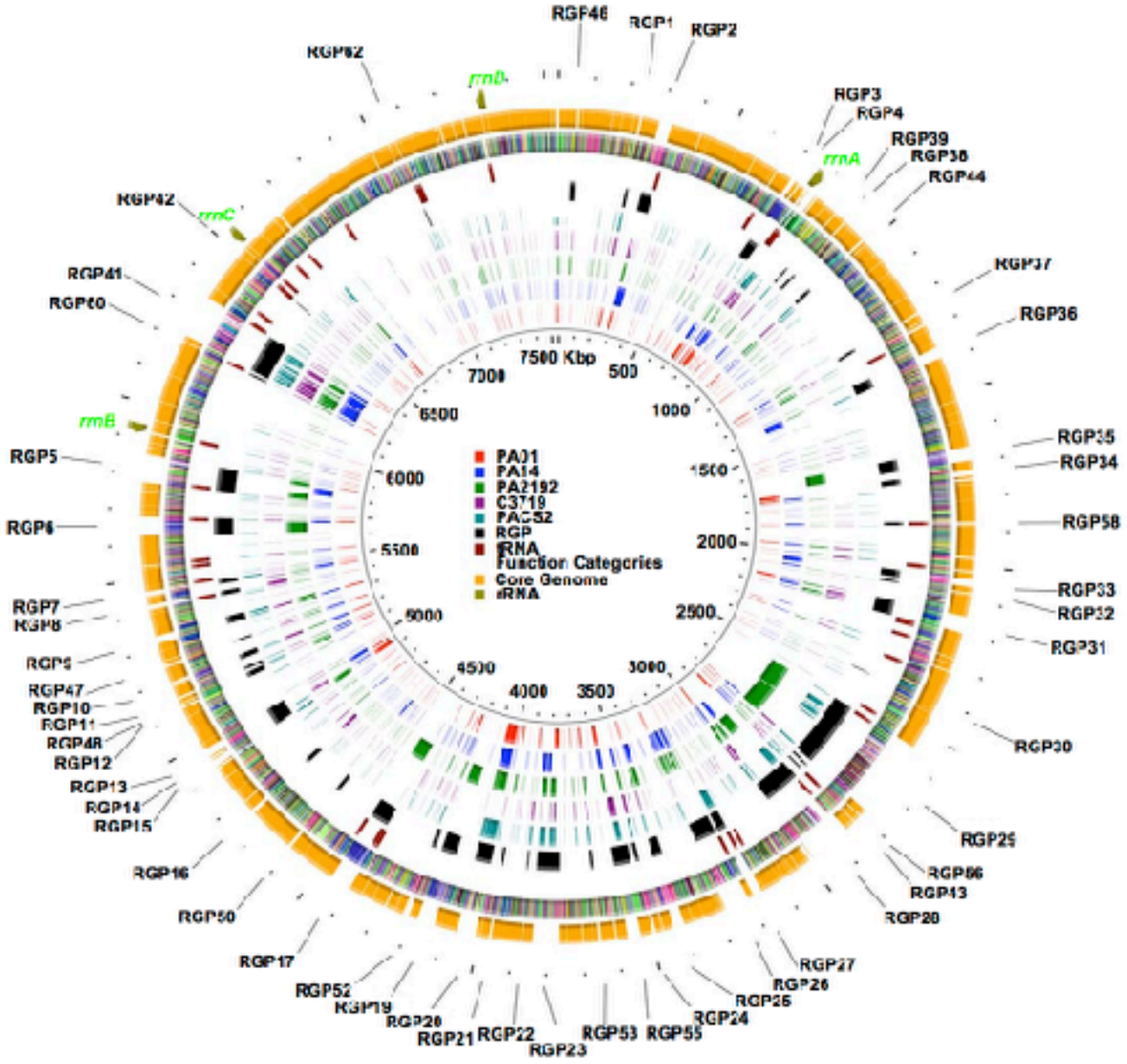
Microarray Analysis

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



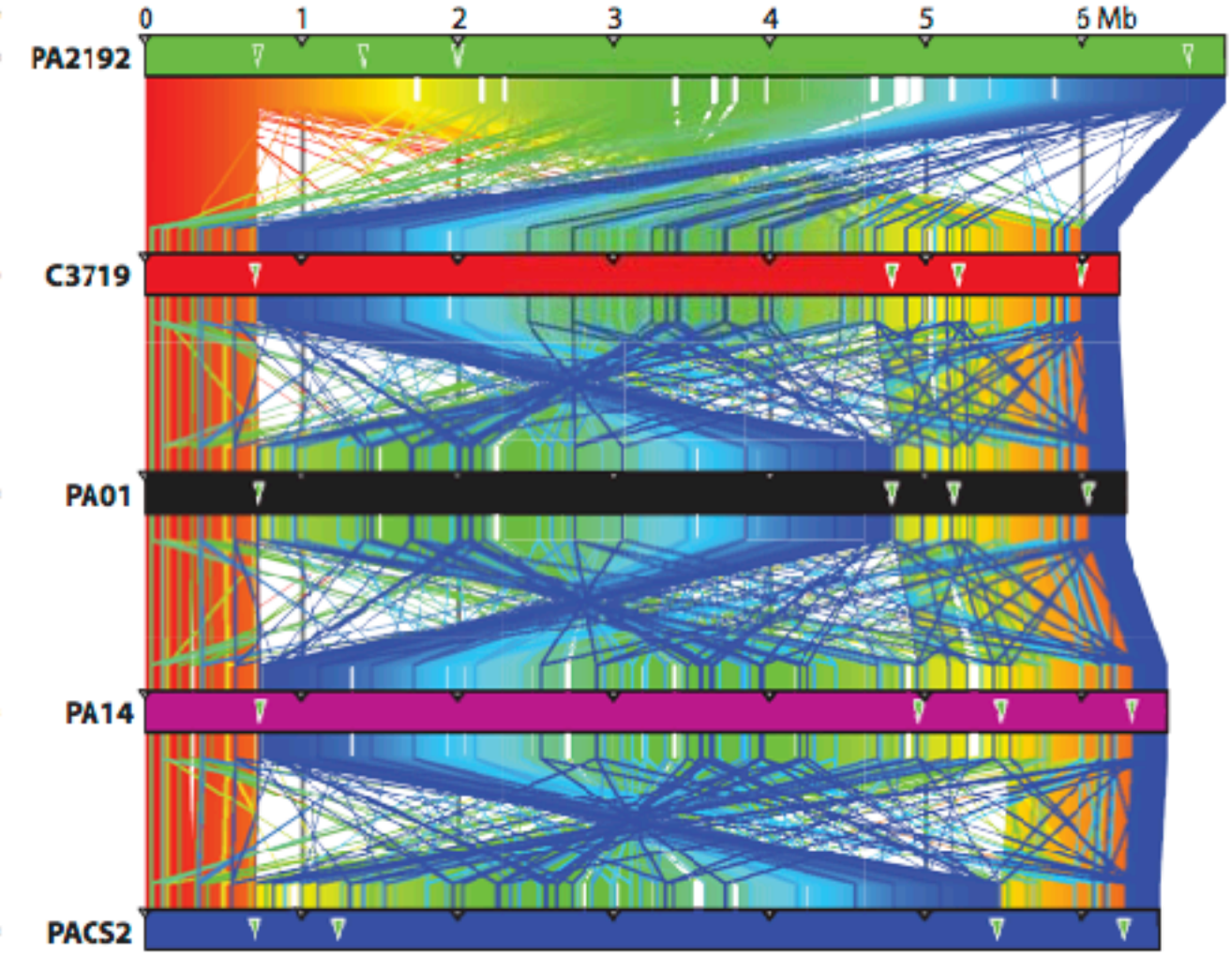
Comparative Genomics

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



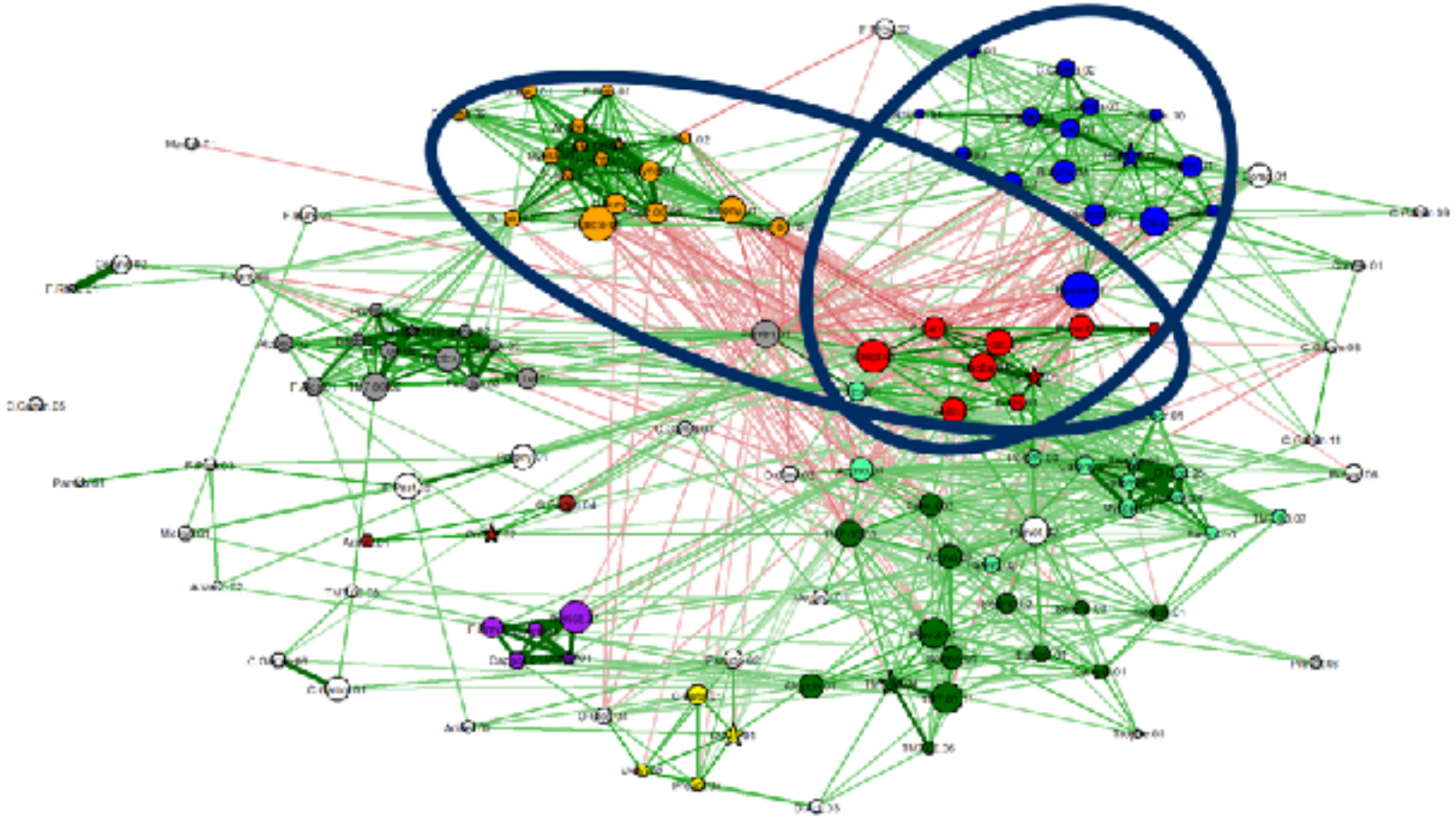
Comparative Genomics

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



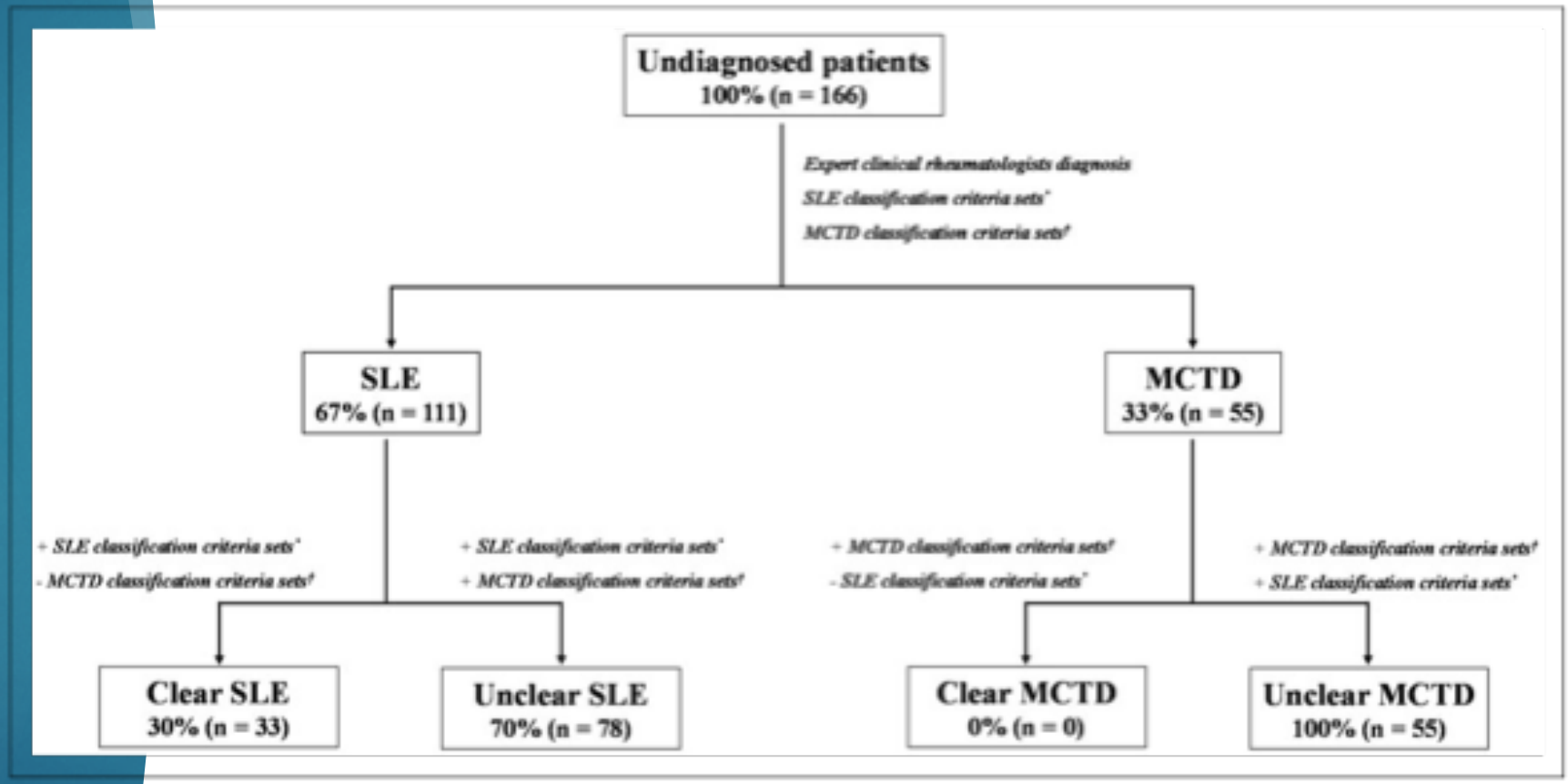
Microbiomes: The Ultimate "Social Network"

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



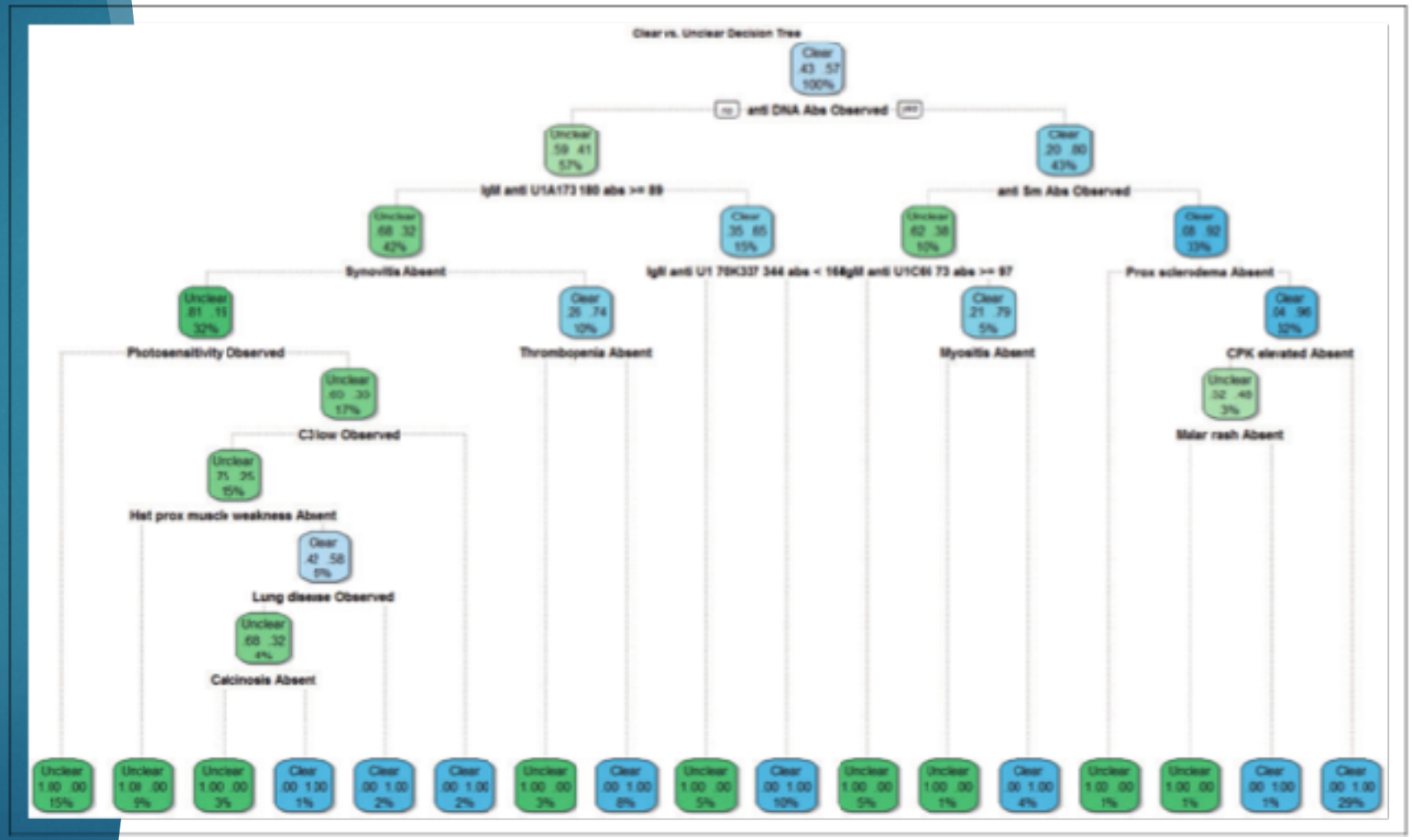
Distinguishing SLE from MCTD

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 …