Bioinformatics Research at FIU

Giri Narasimhan

FLORIDA INTERNATIONAL UNIVERSITY

School of Computer Science

How to do interdisciplinary research?

- Learn the language of the disciplines.
- Learn the culture of the disciplines.
- Find out the problems of the disciplines.
- Understand the limitations of the disciplines.

FAQ: What background do I need in order to study Bioinformatics?

- Depends!
  - Tech Support
  - Creative Problem Solving
- Research in Bioinformatics
  - Appropriate biological skills
  - Quantitative skills (Logic, Probability & Statistics, Discrete Mathematics)
  - Computer Science (Basic programming, internet programming, Algorithms, Databases)

Bioinformatics Research Group (BioRG)

- Pattern Discovery
  - Biomolecular Sequence Data
  - Biomolecular Structure Data
- Microarray Data Analysis
- Primer & Probe Design
- Phylogenetic Analysis
- Ecoinformatics
- Protein Engineering
- Image Processing
  - Biofilm Images
  - Medical Images

Why Pattern Discovery?

- Modern Biomedical Research
  - Generate a “ton of data” and hope that the patterns jump at you!
- Pattern Discovery facilitates this process!

Pattern Discovery in Bioinformatics

- Pattern Discovery in sequences
- Pattern Discovery in structures
- Pattern Discovery in quantitative data

Biomolecular Sequences: Basics

- DNA – 4-letter alphabet **{A, C, G, T}**
- RNA – 4-letter alphabet **{A, C, G, U}**
- Protein – 20-letter alphabet
  {A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, U, V, Y }

HTH Motif Detection

- MTDKMQSLALAPVGNLDSYIRAANAWPMLSADEERALAEKLHYHGDLEAA
  KTLILSHLRFVVHIARNYAGYGLPQADLIQEGNIGLMKAVRRFNPEVGVR
  LVSFAVHWIKAEIHEYVLRNWRIVKVATTKAQRKLFFNLRKTKQRLGWFN
  QDEVEMVARELGVTSKDVREMESRMAAQDMTFDLSSDDDSDSQPMAPVLY
  LQDKSSNFADGIEDDNWEEQAANRLTDAMQGLDERSQDIIRARWLDEDNK
  STLQELADRYGVSAERVRQLEKNAMKKLRAAIEA

- MTDKMQSLALAPVGNLDSYIRAANAWPMLSADEERALAEKLHYHGDLEAA
  KTLILSHLRFVVHIARNYAGYGLPQADLIQEGNIGLMKAVRRFNPEVGVR
  LVSFAVHWIKAEIHEYVLRNWRIVKVATTKAQRKLFFNLRKTKQRLGWFN
  QDEVEMVARELGVTSKDVREMESRMAAQDMTFDLSSDDDSDSQPMAPVLY
  LQDKSSNFADGIEDDNWEEQAANRLTDAMQGLDERSQDIIRARWLDEDNK
  STLQELADRYGVSAERVRQLEKNAMKKLRAAIEA

Collaborator: Kalai Mathee

Start with known examples (Training)

Loc	*Protein*	*Helix 2*									*Turn*				*Helix 3*								
	Name	-1	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
14	**Cro**	F	G	Q	E	K	T	A	K	D	L	G	V	Y	Q	S	A	I	N	K	A	I	H
16	**434 Cro**	M	T	Q	T	E	L	A	T	K	A	G	V	K	Q	Q	S	I	Q	L	I	E	A
11	**P22 Cro**	G	T	Q	R	A	V	A	K	A	L	G	I	S	D	A	A	V	S	Q	W	K	E
31	**Rep**	L	S	Q	E	S	V	A	D	K	M	G	M	G	Q	S	G	V	G	A	L	F	N
16	**434 Rep**	L	N	Q	A	E	L	A	Q	K	V	G	T	T	Q	Q	S	I	E	Q	L	E	N
19	**P22 Rep**	I	R	Q	A	A	L	G	K	M	V	G	V	S	N	V	A	I	S	Q	W	E	R
24	**CII**	L	G	T	E	K	T	A	E	A	V	G	V	D	K	S	Q	I	S	R	W	K	R
4	**LacR**	V	T	L	Y	D	V	A	E	Y	A	G	V	S	Y	Q	T	V	S	R	V	V	N
167	**CAP**	I	T	R	Q	E	I	G	Q	I	V	G	C	S	R	E	T	V	G	R	I	L	K
66	**TrpR**	M	S	Q	R	E	L	K	N	E	L	G	A	G	I	A	T	I	T	R	G	S	N
22	**BlaA Pv**	L	N	F	T	K	A	A	L	E	L	Y	V	T	Q	G	A	V	S	Q	Q	V	R
23	**TrpI Ps**	N	S	V	S	Q	A	A	E	Q	L	H	V	T	H	G	A	V	S	R	Q	L	K

Sequence Patterns

Loc	*Protein*	*Helix 2*									*Turn*				*Helix 3*								
	Name	-1	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
14	**Cro**	F	G	Q	E	K	T	A	K	D	L	G	V	Y	Q	S	A	I	N	K	A	I	H
16	**434 Cro**	M	T	Q	T	E	L	A	T	K	A	G	V	K	Q	Q	S	I	Q	L	I	E	A
11	**P22 Cro**	G	T	Q	R	A	V	A	K	A	L	G	I	S	D	A	A	V	S	Q	W	K	E
31	**Rep**	L	S	Q	E	S	V	A	D	K	M	G	M	G	Q	S	G	V	G	A	L	F	N
16	**434 Rep**	L	N	Q	A	E	L	A	Q	K	V	G	T	T	Q	Q	S	I	E	Q	L	E	N
19	**P22 Rep**	I	R	Q	A	A	L	G	K	M	V	G	V	S	N	V	A	I	S	Q	W	E	R
24	**CII**	L	G	T	E	K	T	A	E	A	V	G	V	D	K	S	Q	I	S	R	W	K	R
4	**LacR**	V	T	L	Y	D	V	A	E	Y	A	G	V	S	Y	Q	T	V	S	R	V	V	N
167	**CAP**	I	T	R	Q	E	I	G	Q	I	V	G	C	S	R	E	T	V	G	R	I	L	K
66	**TrpR**	M	S	Q	R	E	L	K	N	E	L	G	A	G	I	A	T	I	T	R	G	S	N
22	**BlaA Pv**	L	N	F	T	K	A	A	L	E	L	Y	V	T	Q	G	A	V	S	Q	Q	V	R
23	**TrpI Ps**	N	S	V	S	Q	A	A	E	Q	L	H	V	T	H	G	A	V	S	R	Q	L	K

- Q1 G9 N20
- A5 G9 V10 I15

Motif Detection Algorithm: GYM

http://www.cs.fiu.edu/~giri/bioinf/GYM/welcome.html

- G. Narasimhan, K. Mathee, Detection of DNA-binding HTH Motifs in Proteins using the Pattern Dictionary Method, *Methods in Enzymology*, Vol. 370, In Press.
- G. Narasimhan, C. Bu, Y. Gao, X. Wang, N. Xu, K. Mathee, Mining for Motifs in Protein Sequences, *Journal of Computational Biology*, 9(5):707-720, 2002.
- Y. Gao, K. Mathee, G. Narasimhan, X. Wang, Motif Detection in Protein Sequences, Proc. of the 6th SPIRE Conference, 63-72, 1999.
- Theses:
  - Y. Gao
  - C. Bu
  - N. Xu

Training Set Design

- Training Set Bias?
- Desirable properties in a training set:
  - Avoids “over-representation”
  - “Covers” most of the training set
  - Incorporates “domain-specific” knowledge
- Implementation: Class Project by Yanli Sun and Zhengyue Deng.

Immunology Application

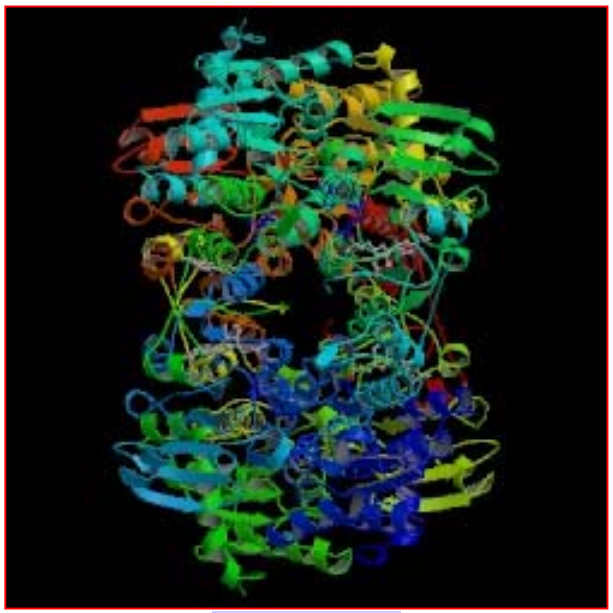
- Pattern Discovery on anti-DNA antibodies from Lupus patients.
- Accurate predictions for one family of antibodies (J558).

Meera Krishnan (2002), Marko Radic

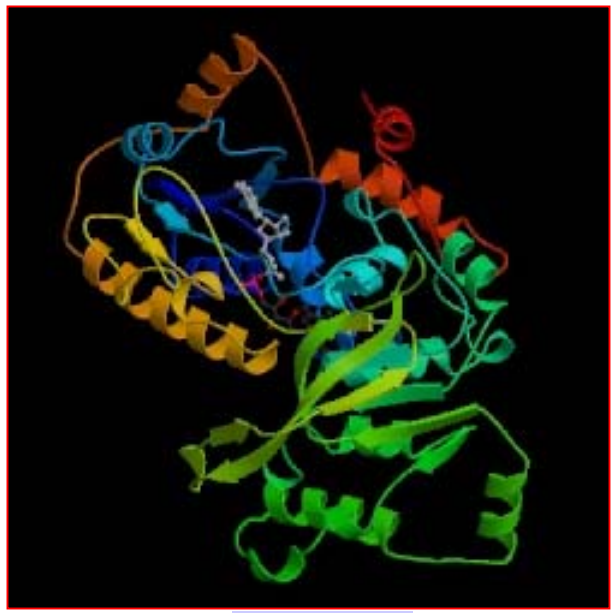
Student Participation

- Yuan Gao (MS 1997, PhD 2001)
- Mu Yang
- Xuning Wang
- Changsong Bu (MS 1999)
- Ning Xu (MS 2000)
- Xiao-rui He
- Junmin Liu
- Meera Krishnan (MS 2001)
- Tom Milledge
- Gaolin Zheng
- Yanli Sun
- Zhengyue Deng

Protein Structures



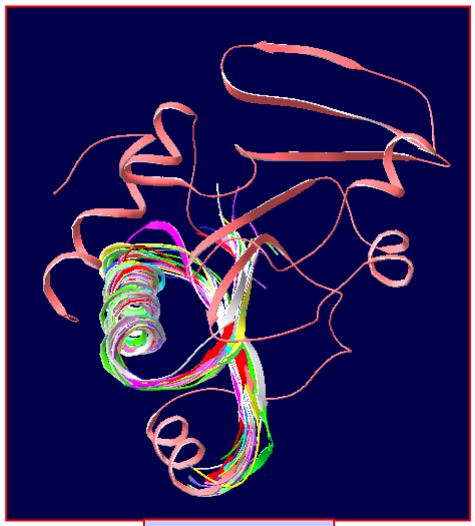
1B3R



1CJC

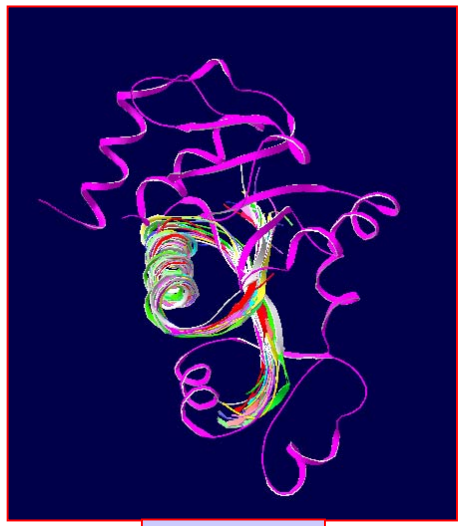
1CF2

Structure Patterns



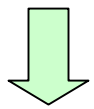
1B3R

1CJC



1CF2

[VILM]-x(2)-G-x(2)-[AGS]-x(3,4)-A-x(1,2)-[VILM]-x(5,7)-[VILM]-x-[VILM]-x-[VILM]-x-[DE]



[VILM]-x-[VILM]-x(3,4)-G-x(2)-[AGS]-x(3,4)-A-x(1,2)-[VILM]-x(6,8)-[VILM]-x-[VILM]-x-[ADEGKNQRST]

Sequence-Structure Patterns (SSP)

- **Automatically Generate SSPs**

Tom Milledge

Chengyong Yang

- **How to store and retrieve SSPs**
  - **Geometric Hashing**

Min Chi Hu

Gene Expression

- Different cells exhibit different gene expressions.

DNA code for Proteins

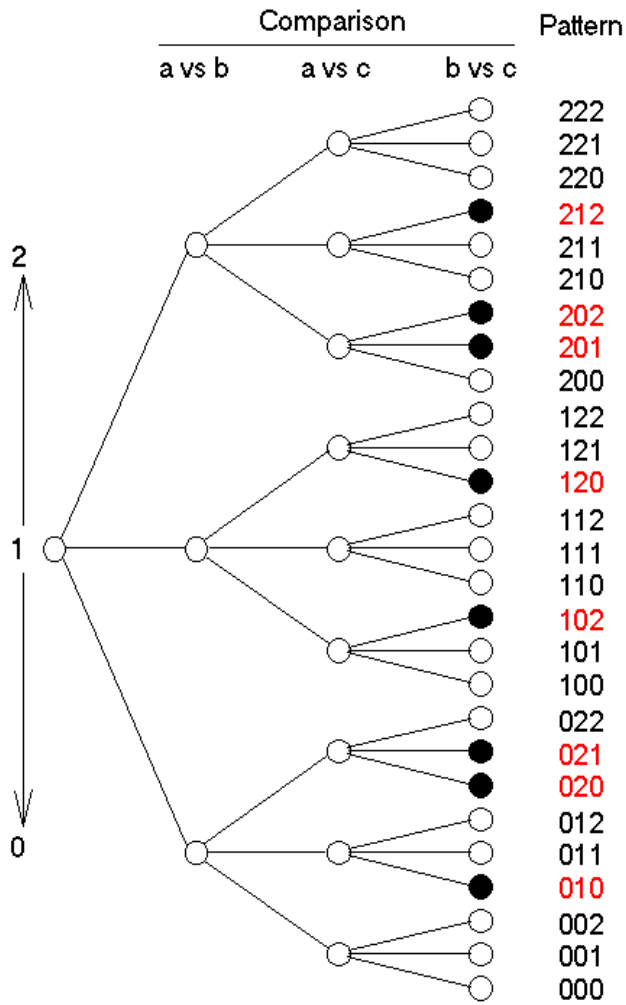
Proteins perform some of life's most essential functions, often working in groups.

Proteins:
Hemoglobin,
Immunoglobin,
Insulin,
Keratin,
Melanin,
Hormones,
Enzymes,
etc.

Microarray Data Analysis

	Before Treatment	After Treatment 1	After Treatment 2	…	After Treatment k
Gene 1	B_1	A_{11}	A_{21}		A_{k1}
Gene 2	B_2	A_{12}	A_{22}		A_{k2}
Gene 3	B_3	A_{13}	A_{23}		A_{k3}
…					
Gene n	B_n	A_{1n}	A_{2n}		A_{kn}

Patterns and Pattern Tree



- Xiao-rui He (2001)
- Peter Dimitrov (2000)
- Software **μ-NP**

Gaolin Zheng

T. Sutter, Xiao-rui He, Peter Dimitrov, L. Xu, G. Narasimhan, E.O.George, *et al.*, Multiple comparisons model-based clustering and ternary pattern tree numerical display of gene response to treatment: Procedure and application to the preclinical evaluation of chemopreventive agents, *Molecular Cancer Therapeutics*, 1(14):1283-1292, 2002.

Microarray Data Analysis - II

	Normal Patient	Cancer Patient
Gene 1	B_1	A_1
Gene 2	B_2	A_2
Gene 3	B_3	A_3
…		
Gene n	B_n	A_n

Classification Using Microarray Data

- Neural Network Classifiers
  - Bayesian methods
  - Bagging
  - Boosting
- Gene Selection

Gaolin Zheng

G. Zheng, G. Narasimhan, E.O. George, Neural Network Classifiers and Gene Selection Methods for Microarray Data on Human Lung Adenocarcinoma, *CAMDA'03*, Nov. 2003.

New Questions

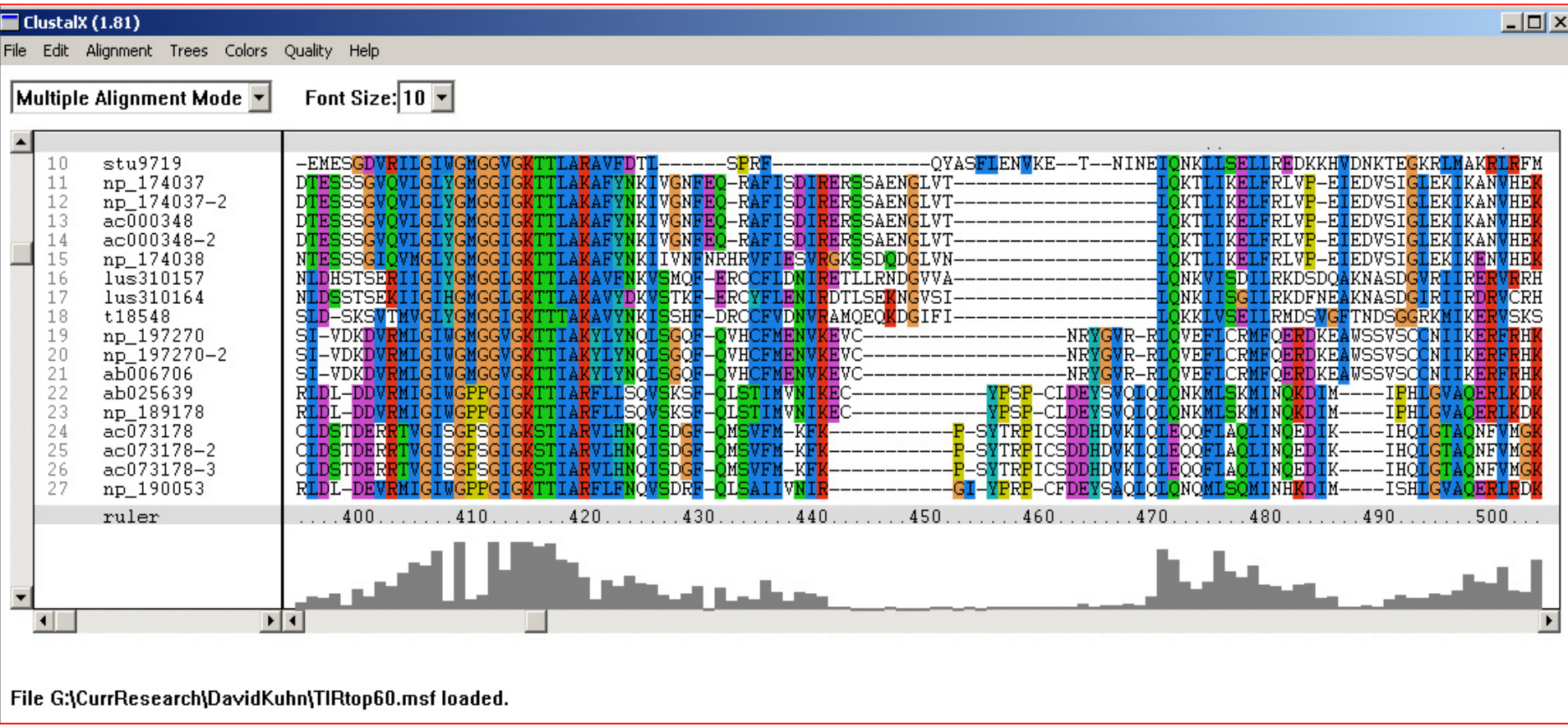
- How to integrate similar data from different laboratories?
- How to find a small set of “significant” genes?
- How to detect significant genes with low expression levels?

- Gaolin Zheng
- Erliang Zeng

Degenerate Primer Design

- Find R-genes in cacao. [David Kuhn]
- Design primers based on known R-genes from other plants.
- Find common substrings in known R-genes.

Find common substring

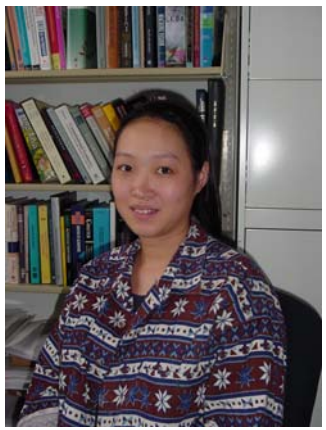


- If common substring cannot be found, make smaller groups.

Degenerate Primer Design

- Find R-genes in cacao. [David Kuhn]
- Design primers based on known R-genes from other plants.
- Find common substrings in known R-genes.

Xintao Wei



Xintao Wei, David Kuhn, G. Narasimhan, Degenerate Primer Design via Clustering, *Proc. Of CSB 2003*, 75-83, August 2003.

Jason Stein, Chris Archer, Jordan Farrow

Probe Design

- How to differentiate between a set of microbes?
- Find short sequences that differentiate all of them.

Daniel Cazalis

Tom Milledge

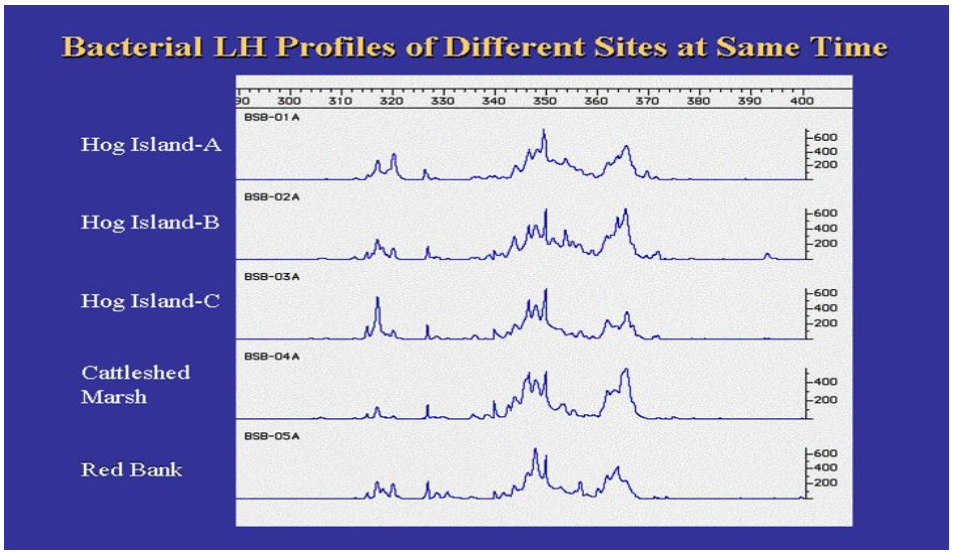
Probe Design

- How to detect the presence of unicellular cyanobacteria (*Synechococcus* sp.) from other multicellular cyanobacteria (*Trichodesmium* sp.)? [Frank Jochem, Kalai Mathee]
- Find sequences that are present in all *Synechococcus* strains and that are not present in *Trichodesmium* strains.

Jason Stein

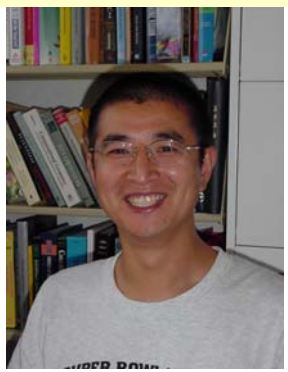
Eco-informatics

- How to differentiate between soil samples? [Dee Mills, Krish Jayachandran, Kalai Mathee]
- Find profiles of microbial communities in them.
  - Unsupervised Clustering
  - Supervised Clustering
- Tested on soil samples from Idaho



Yong Wang

Chengyong Yang

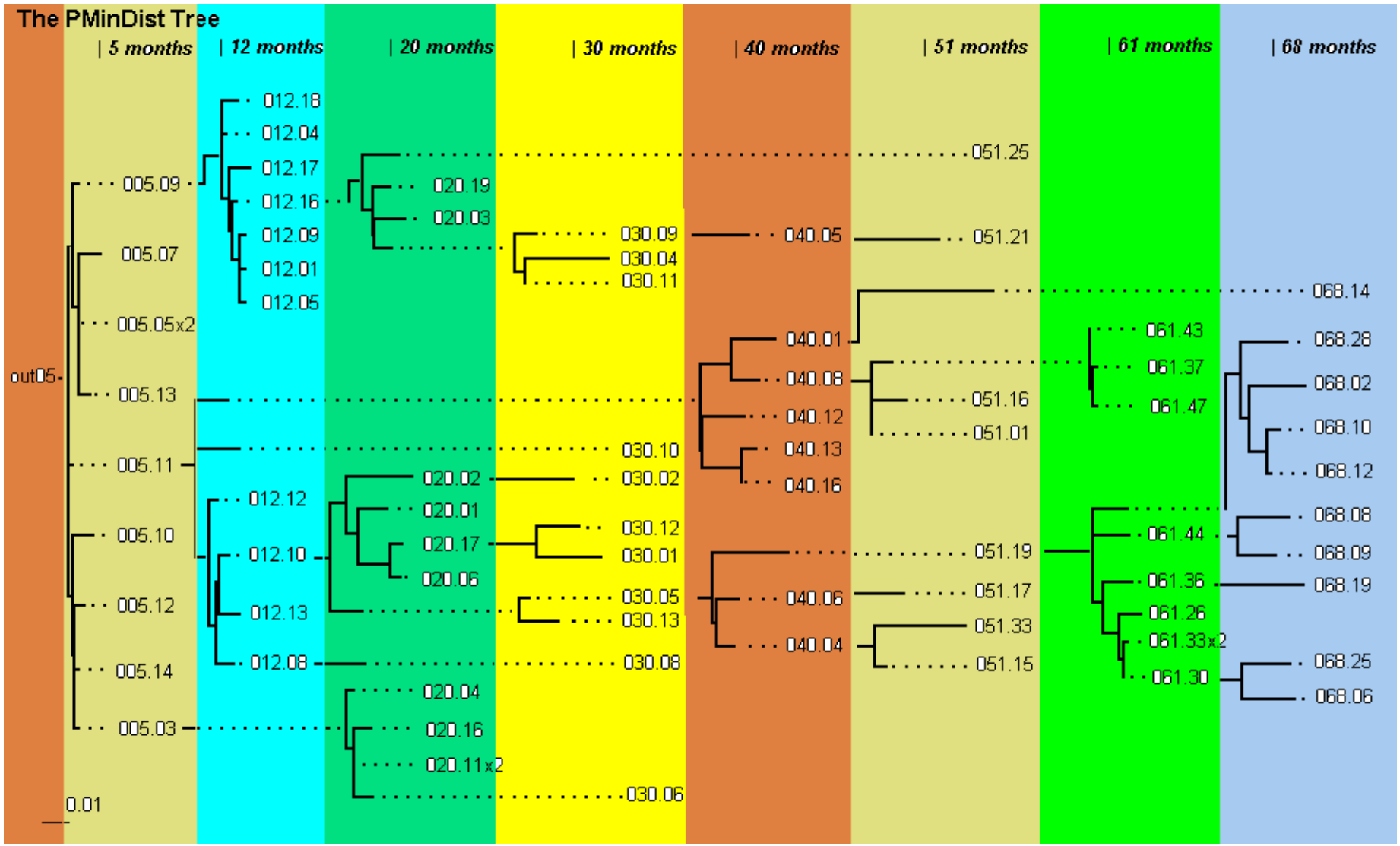


HIV Sequence Analysis

- Serially Sampled Quasispecies
- Given HIV sequences from a single patient sampled at different times, how to analyze these sequences?

Patricia Buendia

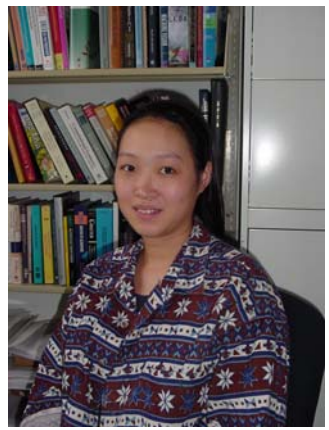
HIV Evolution Tree



Protein Engineering

- How to design bacteriorhodopsins with higher thermal stability? [Renugopalakrishnan]
- Look at similar proteins from thermophiles and mesophiles and study their differences in a systematic manner. [Ferredoxins]

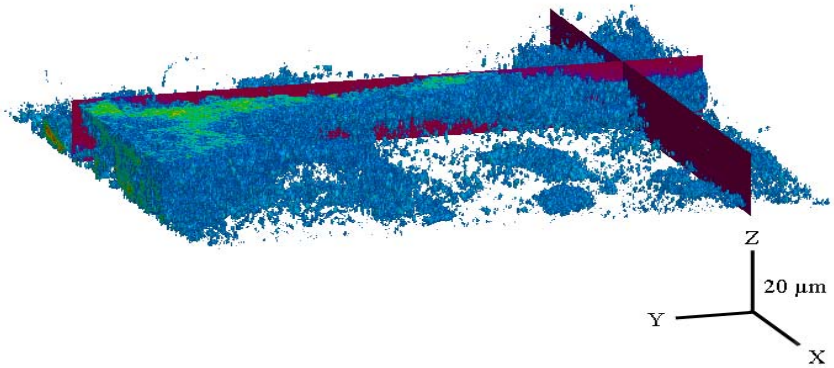
Xintao Wei

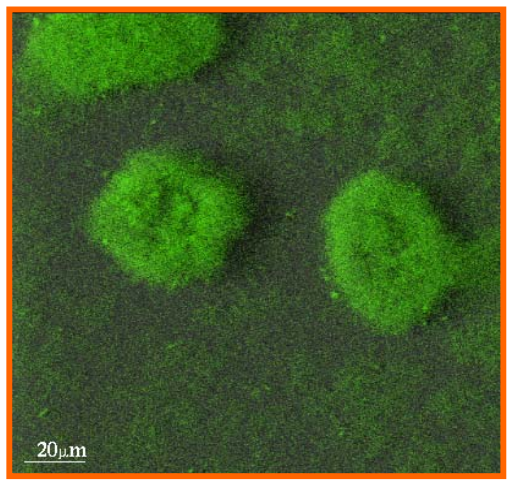


Chem-Informatics

- Cassian D’Cunha [David Chatfield]

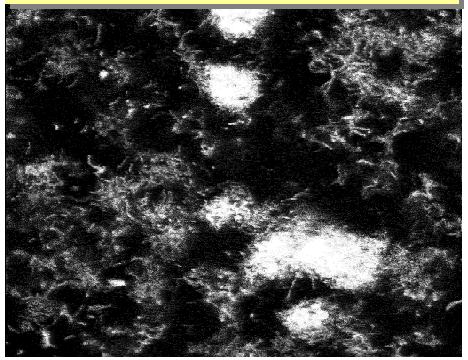
Biofilm Image Processing



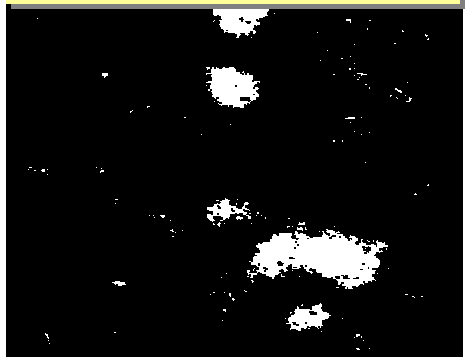


Kalai Mathee, Søren Molin

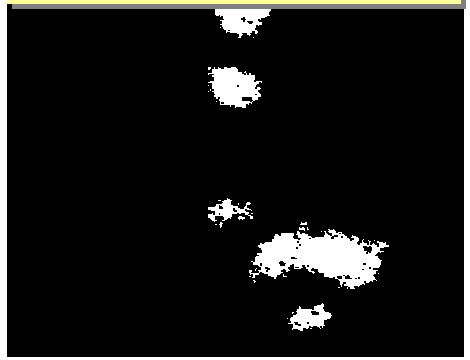
Original CLSM Image



Threshold image

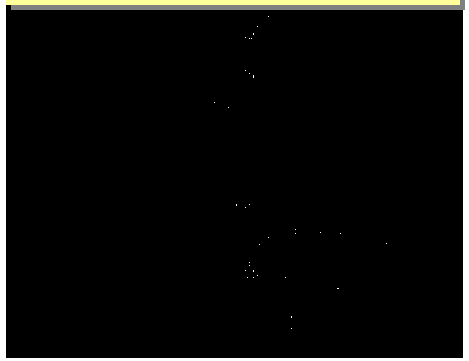


Delete small islands

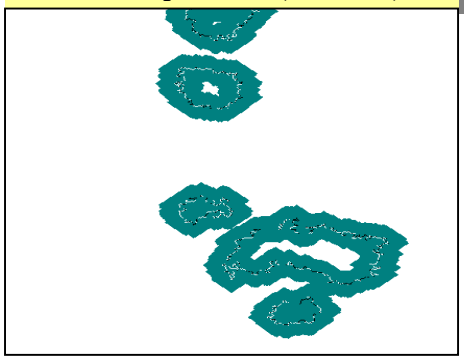


Fill out small lakes

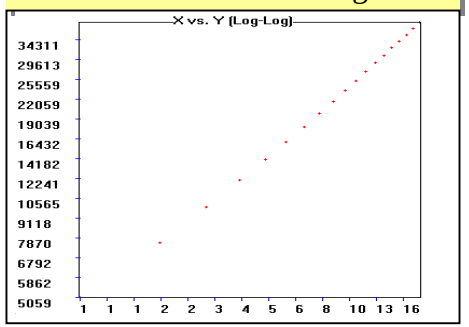
Detect edges



FD computation (Dilation)



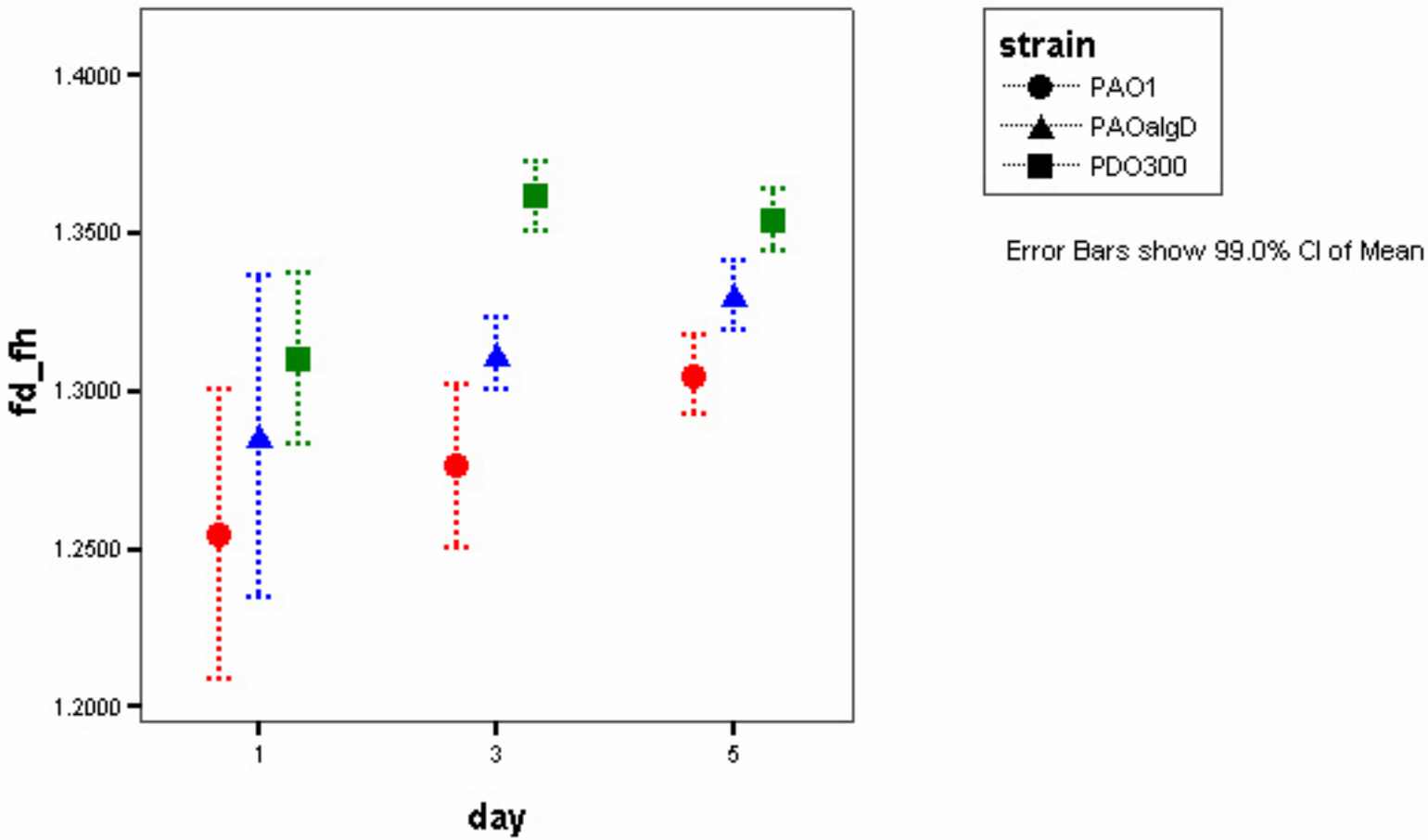
Plot Perimeter vs Length



BIP

Fractal Dimension Values vs. Day

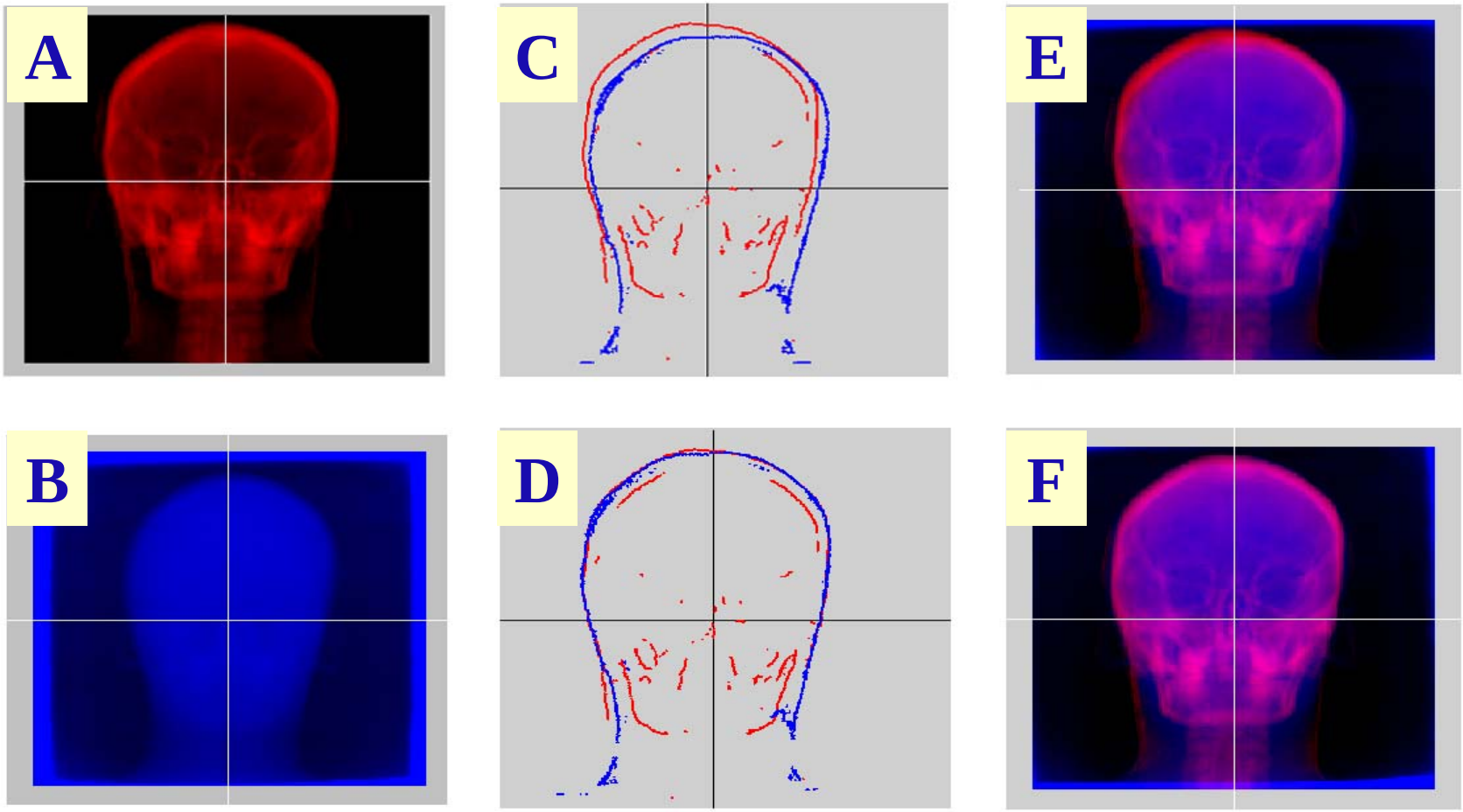
FD by Fast Hybrid method (fd_fh) vs. Day (day)



Biofilm Image Processing

- BIP Software
  - Qichang Li
  - Zhou Ji
  - Shalaka Indulkar

Medical Image Registration



Andres Parra, U of Memphis; S. Samant, St. Jude Research Hospital

Bioinformatics Research Group (BioRG)

- Pattern Discovery
  - Sequence Data
  - Structure Data
- Microarray Data Analysis
- Primer & Probe Design
- Phylogenetic Analysis
- Ecoinformatics
- Protein Engineering
- Image Processing
  - Biofilm Images
  - Medical Images

- Unsupervised Structure Pattern Discovery
- Protein Structure Prediction
- Gene Network Prediction
- Phylogenetic Analysis in the presence of recombination
- Informatics: Chemistry, Material Science, Forensics, Finance
- Whole Genome Comparison
- Community Biofilm Analysis

Bioinformatics Research Group (BioRG)

- We are BioRG (pronounced Borg).
- We can assimilate you!

Questions??

GYM 2.0

Contact Person:
Giri Narasimhan

Authors:
Changsong Bu
Giri Narasimhan
Kalai Mathee

Master Set used for training the program:
Master Set

Number of Hits:
678

Detecting Helix-Turn-Helix motifs in Proteins

GYM 2.0 can be used to detect Helix-Turn-Helix Motifs in protein sequences.

When citing this work, please cite the following papers:

- G. Narasimhan, C. Bu, Y. Gao, X. Wang, N. Xu, K. Mathee, Mining for Motifs in Protein Sequences, *Journal of Computational Biology*, **9**(5):707-720, 2002.
- Y. Gao, K. Mathee, G. Narasimhan, X. Wang, Motif Detection in Protein Sequences, Proc. of the 6th SPIRE Conference, 63-72, 1999.

For using this program, please enter the protein sequence in the text area.

If you use the GYM 2.0 program in your research, please cite the papers listed above.

Please copy and paste the sequence in the text area.
(No spaces or end of line characters).

```
MTDKMQSLALAPVGNLDSYIRAANAWPMLSADEERALAEKLHYHGDLEAA□KTLILSHLRFVVHIARNYAGYGLP
QADLIQEGNIGLMKAVRRFNPEVGVR□LVSFAVHWIKAEIHEYVLRNWRIVKVATTKAQRKLFFNLRKTKQRLGW
FN□QDEVEMVARELGVTSKDVREMESRMAAQDMTFDLSSDDDSDSQPMAPVLY□LQDKSSNFADGIEDDNWEEQA
ANRLTDAMQGLDERSQDIIRARWLDEDNK□STLQELADRYGVSAERVRQLEKNAMKKLRAAIEA
```

Submit Reset

GYM 2.0: Results for Your Sequence

*** GYM Results Summary ***

Input Sequence with Highlighted HTH Motif Locations:

MTDKMQSLALAPVGNLDSYIRAANAWPMLSADEERALAEKLHYHGDLEAA
KTLILSHLRFVVHIARNYAGYGLPQADLIQEGNIGLMKAVRRFNPEVGV
R LVSFAVHWIKAEIHEYVLRNWRIVKVATTKAQRKLFFNLRKTKQRLGW
FN QDEVEMVARELGVTSKDVREMESRMAAQDMTFDLSSDDDSDSQPMAP
VLY LQDKSSNFADGIEDDNWEEQAANRLTDAMQGLDERSQDIIRARWLD
EDNK STLQELADRYGVSAERVRQLEKNAMKKLRAAIEA

Length of Sequence = 289

Predicted HTH Motif Locations:

Pick	Loc	LP	NPM	Score	Detected?	Motif
Best	256	5	72	80	+	STLQELADRYGVSAERVRQLEK
2nd	155	4	8	41	+	DEVEMVARELGVTSKDVREMES

---- GYM's BEST MATCH: Details ----

Start Location (Loc): 256

Length of Longest Pattern Matched (LP): 5

Number of Patterns Matched (NPM): 72

Number of Locations Matched: 17

The Maximum blosum score of the whole sequence (Score): 80

Best Motif Sequence:

```
0123456789 0123456789 01
STLQELADRY GVSAERVRQL EK
_^^_^^^^^_ ^^^_^_^^^^ ^^
```

Predicted HTH Motif Locations:

Pick	Loc	LP	NPM	Score	Detected?	Motif
Best	256	5	72	80	+	STLQELADRYGVSAERVRQLEK
2nd	155	4	8	41	+	DEVEMVARELGVTSKDVREMES

---- GYM's BEST MATCH: Details ----

Start Location (Loc): 256
Length of Longest Pattern Matched (LP): 5
Number of Patterns Matched (NPM): 72
Number of Locations Matched: 17
The Maximum blosum score of the whole sequence (Score): 80

Best Motif Sequence:

```
0123456789 0123456789 01
STLQELADRY GVSAERVRQL EK
_^^_^^^^^_ ^^^_^_^^^^ ^^
```

---------- Sequence Classification ----------

```
Group#    Pattern                Proteins
526     | A6 S12 T1 V11 V16 | LacI Ec, CytR Ec, PurR Ec, RbtR Ka,
522     | A6 S12 T1 V16 | LacI Ec, CytR Ec, PurR Ec, Cro P22, SpoIIAC Bs,
518     | A6 G10 S12 T1 V16 | LacI Ec, RbtR Ka, Cro P22, SpoIIAC Bs,
517     | A6 G10 S12 T1 | LacI Ec, Cro P22, SpoIIAC Bs, SpoIIG Bs,
496     | E4 G10 S12 V16 | CAP Ec, Cox P2, SpoIIAC Bs, SigB Bs,
493     | E4 E14 E20 R17 V11 | MerR BR, MerR Sa, rpoD Ec, rpoH Ec, sigA Bs,
491     | A6 S12 T1 V11 | LacI Ec, CytR Ec, PurR Ec, ApL 186,
472     | A6 K21 L19 S12 T1 | Ada Ec, rpoH Ec, SpoIIAC Bs, SpoIIG Bs,
467     | A6 G10 T1 V11 | LacI Ec, MerR Tn21, TetR Tn10, ApL 186,
466     | A6 G10 S12 V11 | LacI Ec, DeoR Ec, AsnC Ec, ApL 186,
465     | A6 G10 S12 T1 V11 | LacI Ec, RbtR Ka, ApL 186, rpoH Ec,
462     | A6 G10 S12 V16 | LacI Ec, Cox P2, SpoIIAC Bs, ParB P1,
461     | A6 G10 S12 T1 | LacI Ec, ApL 186, SpoIIAC Bs, SpoIIG Bs,
460     | A6 G10 S12 T1 | Cro P22, C 16-3 Rm, SpoIIAC Bs, SpoIIG Bs,
```

BioRG Research Interests

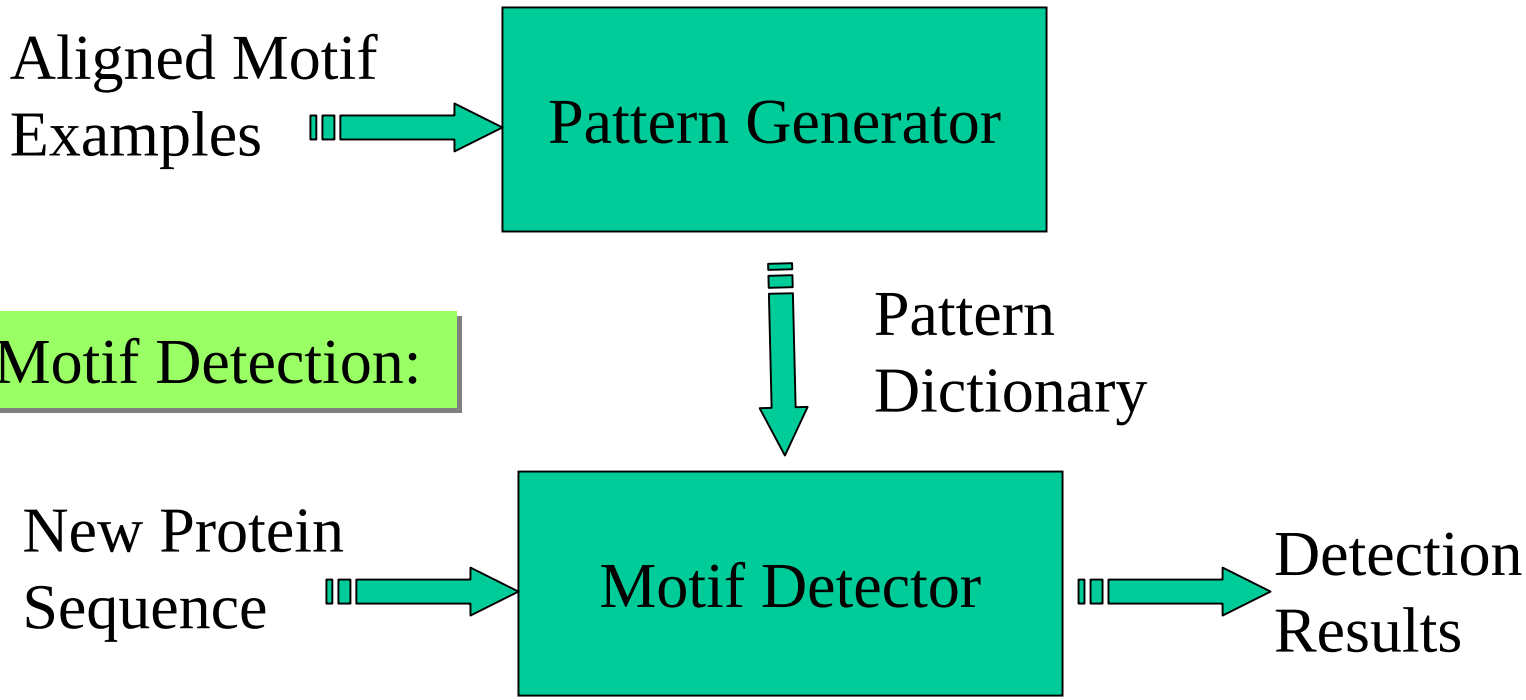
- Pattern Discovery
  - Biomolecular Sequences
  - Biomolecular Structures
- Microarray Data Analysis
- Primer Design
- Probe Design
- Ecoinformatics
- Protein Engineering
- Image Processing
  - Biofilm Images
  - Medical Images
- Phylogenetics: HIV Evolution





Motif Detection Algorithm: GYM

Pattern Generation:



http://www.cs.fiu.edu/~giri/bioinf/GYM/welcome.html

Why Pattern Discovery?

- Axiom: Want to understand the world around us.
- The "Book of Rules" is missing!
- Empirical Science
  - Observe the world around and discern the "patterns".
- Modern Science
  - Generate a "ton of data" and wait for the patterns to jump at you!
- Pattern Discovery facilitates this process!