What is Bioinformatics? Why Bioinformatics?

Giri Narasimhan
Florida International University
Miami, FL 33199, USA.

Molecular Biology Primer

Organisms | Cells | Chromosomes | DNA

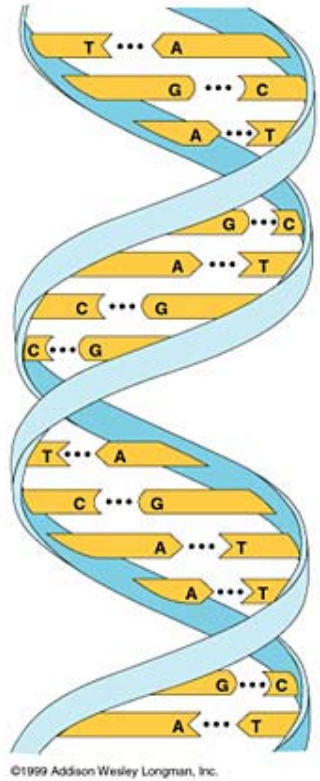
DNA = Chain of Nucleotides

(Two strands; double helix)

...TTCTGCATTCGGTGAAGAGGCGCTCTAG...

...AAGACGTAAGCCACTTCTCCGCGAGATC...

Genome



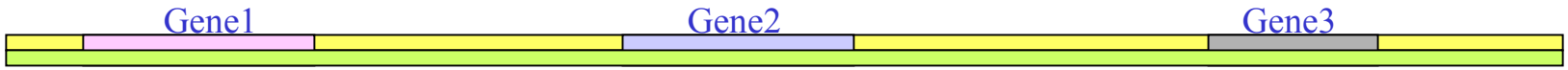
Molecular Biology Primer (Cont'd)

DNA code for Proteins

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

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

Central Dogma



DNA

Transcription

mRNA

Translation

Protein

Genetic Code

Nucleotides in DNA (4) : **A,G,T,C**

Amino Acids in Protein (20):

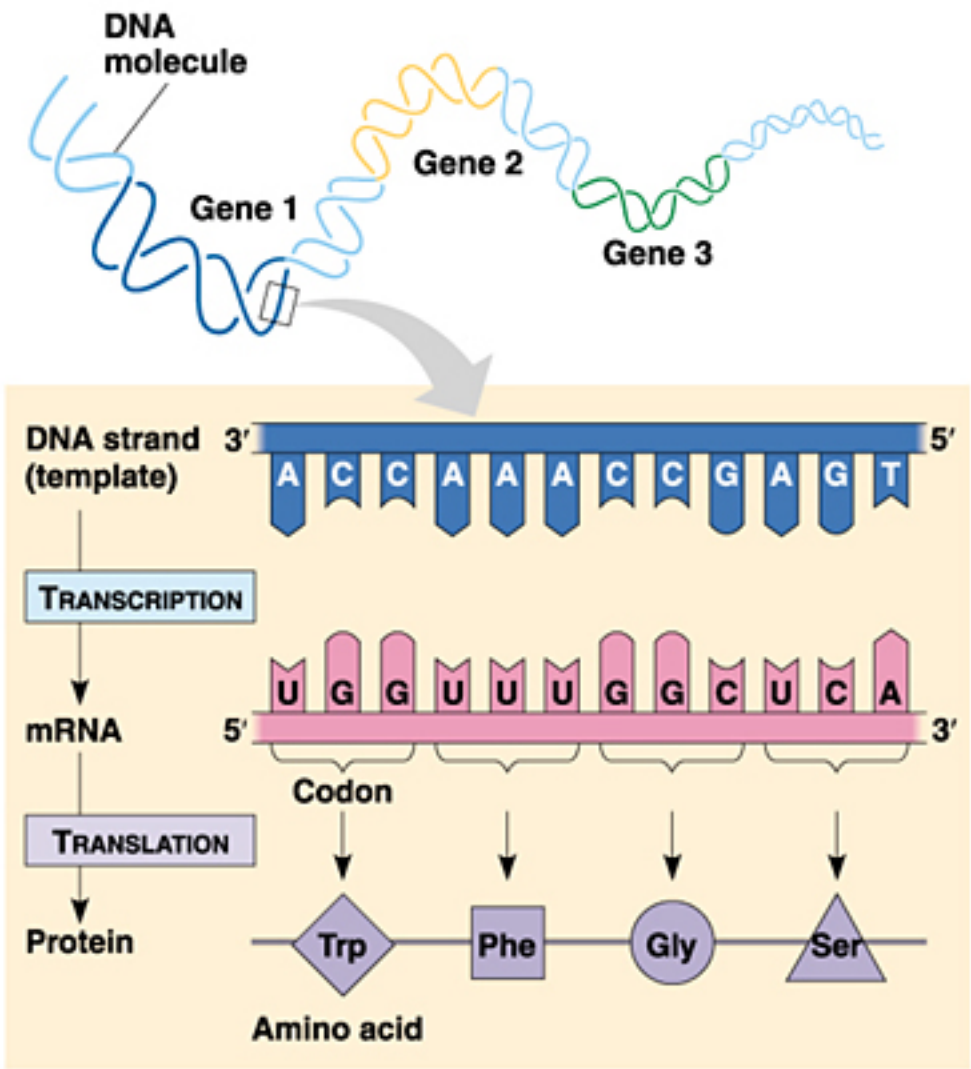
A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y

GAT	TCG	ATG	GCG	CCT	GTA
D	S	M	A	C	V

Nucleotides in RNA (4) : **A,G,U,C**

Triplet Code

- one gene = one protein



The Genetic Code – Triplet Code

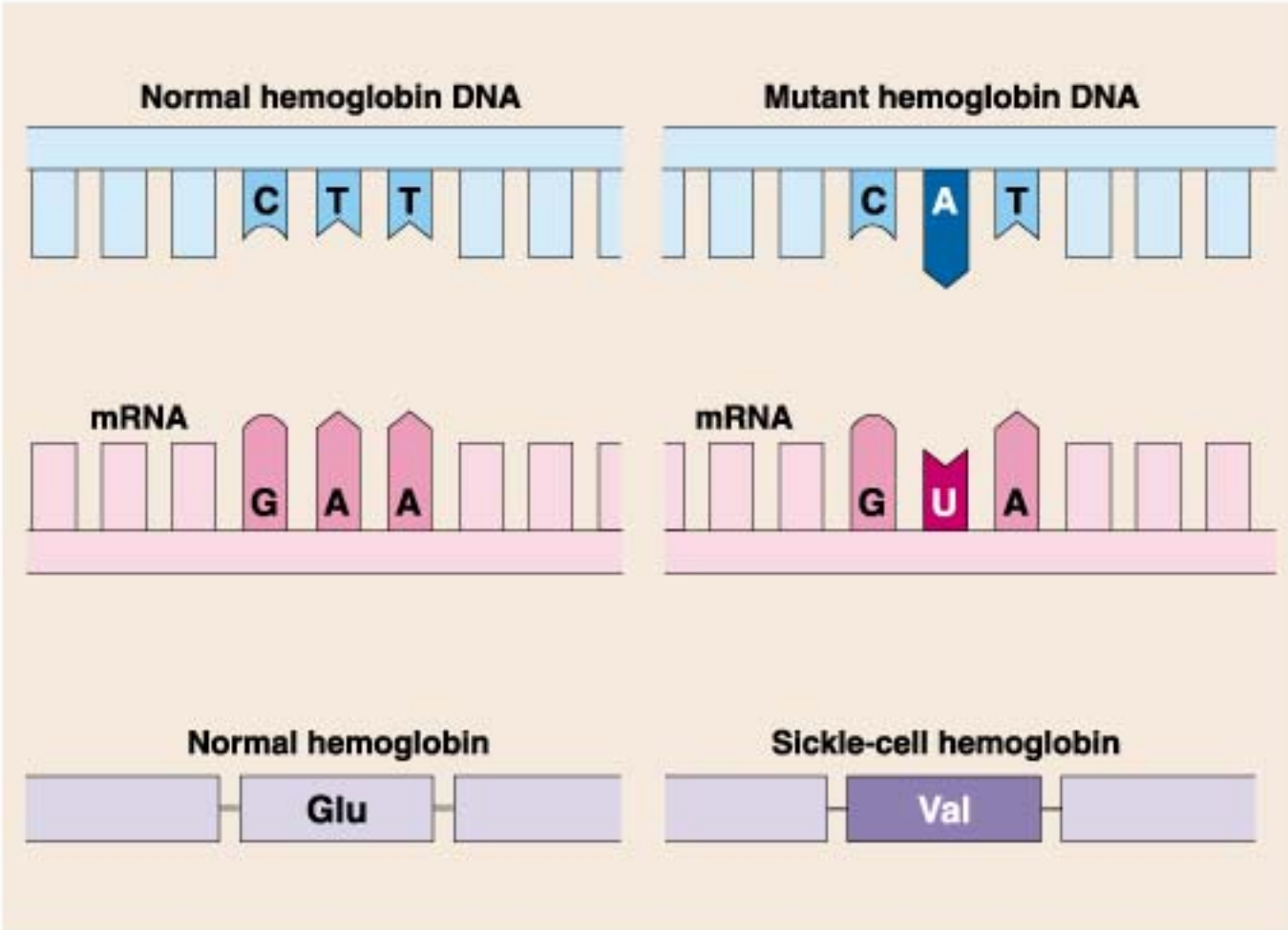
1st position (5' end)	2nd position U	2nd position C	2nd position A	2nd position G	3rd position (3' end)
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Degeneracy in Code

Mutations in Genes

- Mutations cause variations – good and bad.
- Mutations are the cause of evolution.
- Mutations in the nucleotides
  - Transitions:
    - C ↔ T or A ↔ G
  - Transversions:
    - A/G ↔ C/T
- Mutations are the cause of diseases
  - Hemophilia
  - Cystic Fibrosis

Sickle Cell Disease

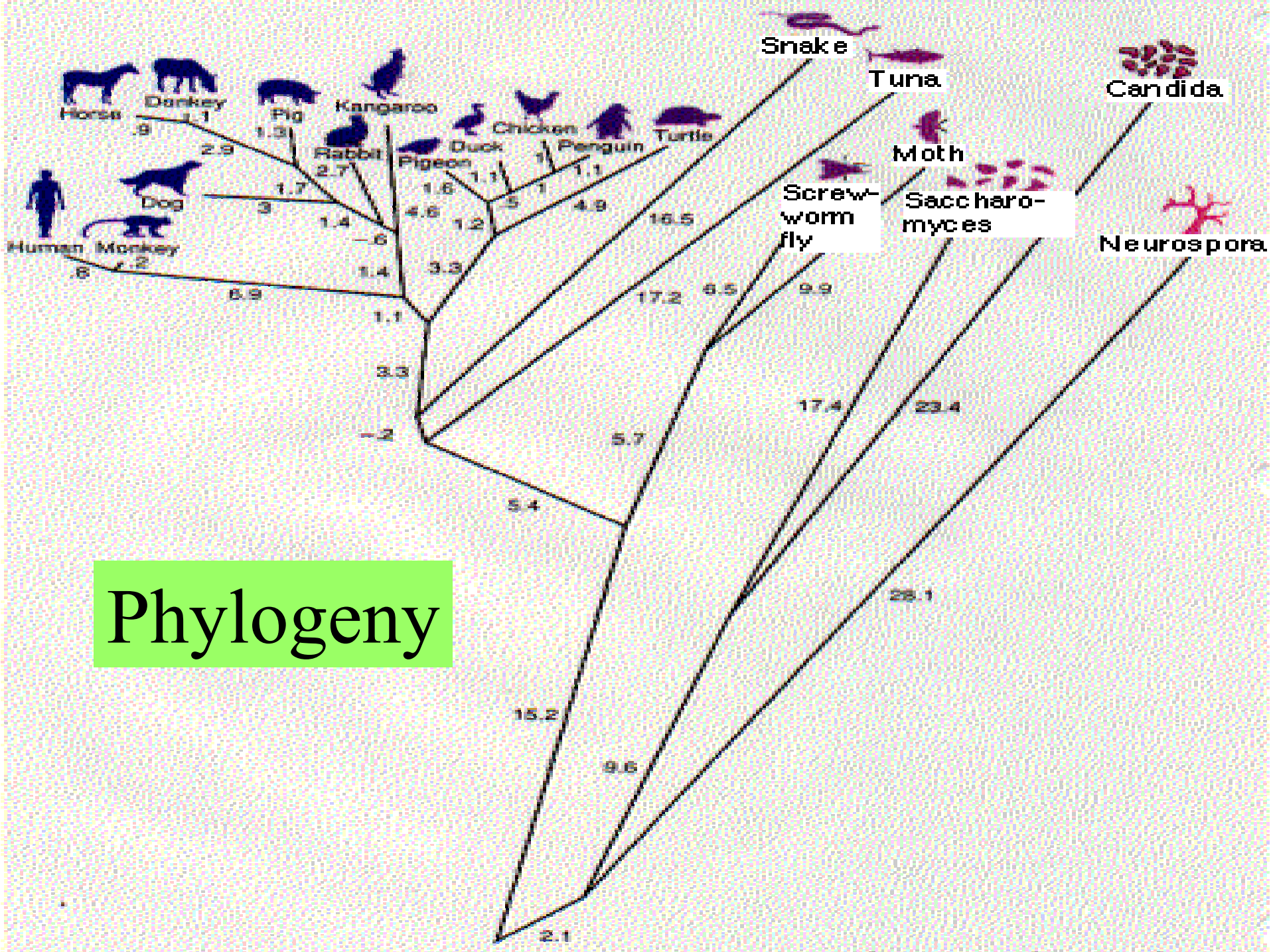


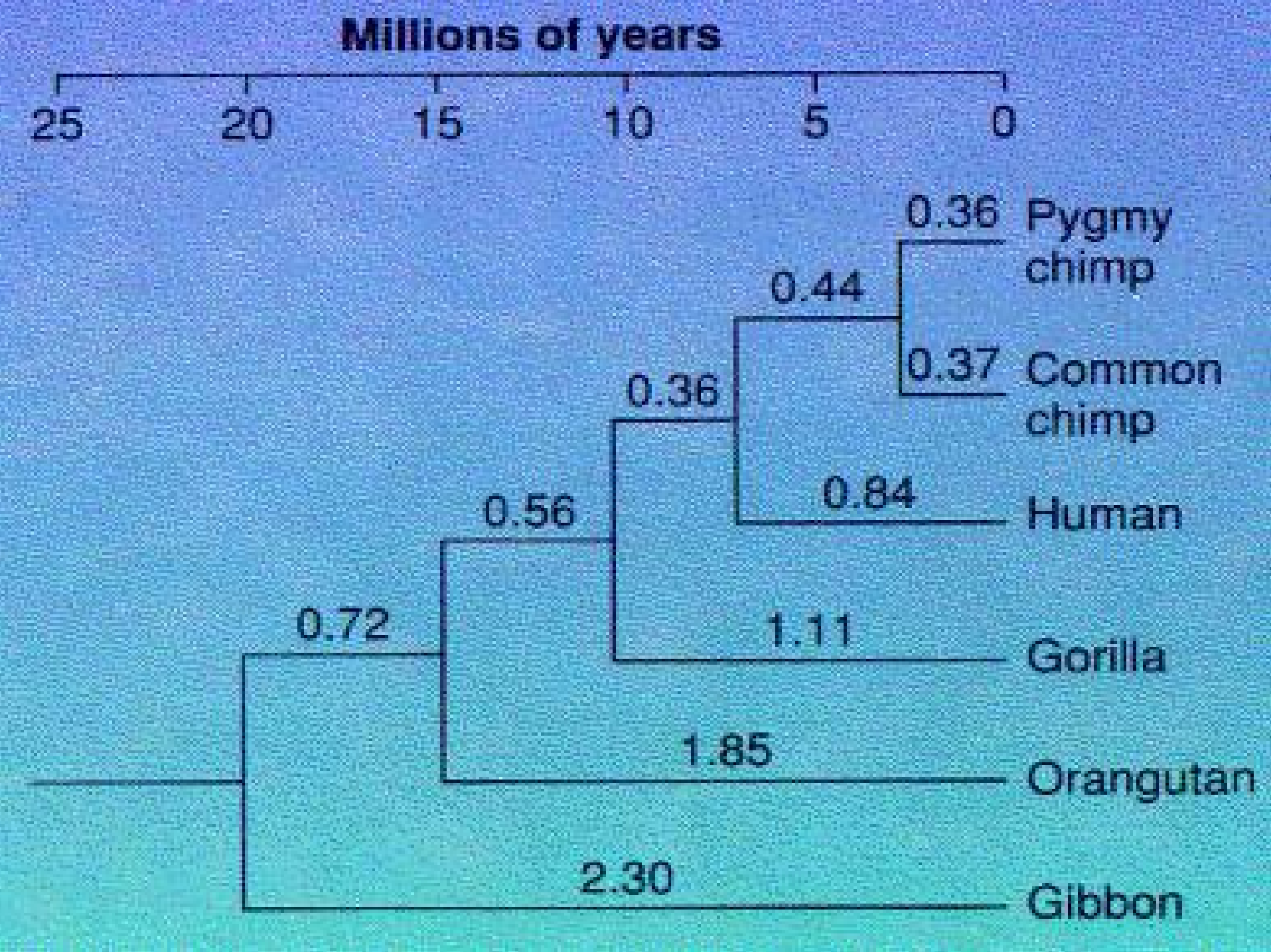
Sequence Alignment

```
Query: 154 GTRVRAMAIYKQSQHMTEVVRRCPHHE--RCSDSDGLAPPQHLIRVEGNLRVEYLDDRNT 211
           G  +RAM +YK+++H+TEVV+RCP+HE  R  +   +APP HLIRVEGN   +Y++D  T
Sbjct: 128 GAVIRAMPVYKKAEHVTEVVKRCPNHELSREFNEGQIAPPSHLIRVEGNSHAQYVEDPIT 187


Query: 212 FRHSVVVPYEPPEVGSDCTTIHYNYMCNSSCMGGMNRRPILTIITLEDSSGNLLGRNSFE 271
            R SV+VPYEPP+VG++ TT+ YN+MCNSSC+GGMNRRPIL I+TLE   G +LGR  FE
Sbjct: 188 GRQSVLVPYEPPQVGTEFTTVLYNFMCNSSCVGGMNRRPILIIVTLETRDGQVLGRRCFE 247
…
```

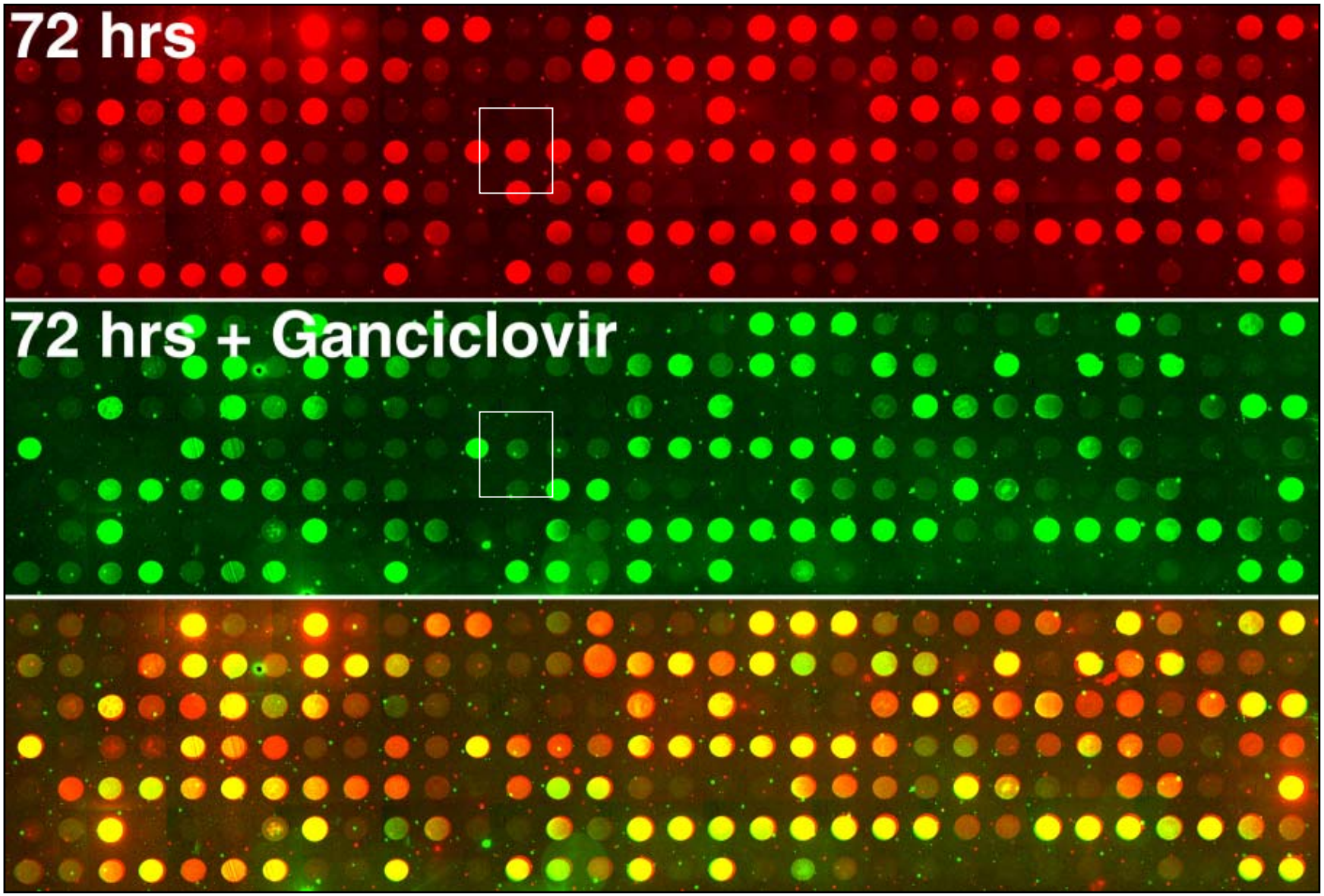
Multiple Sequence Alignment

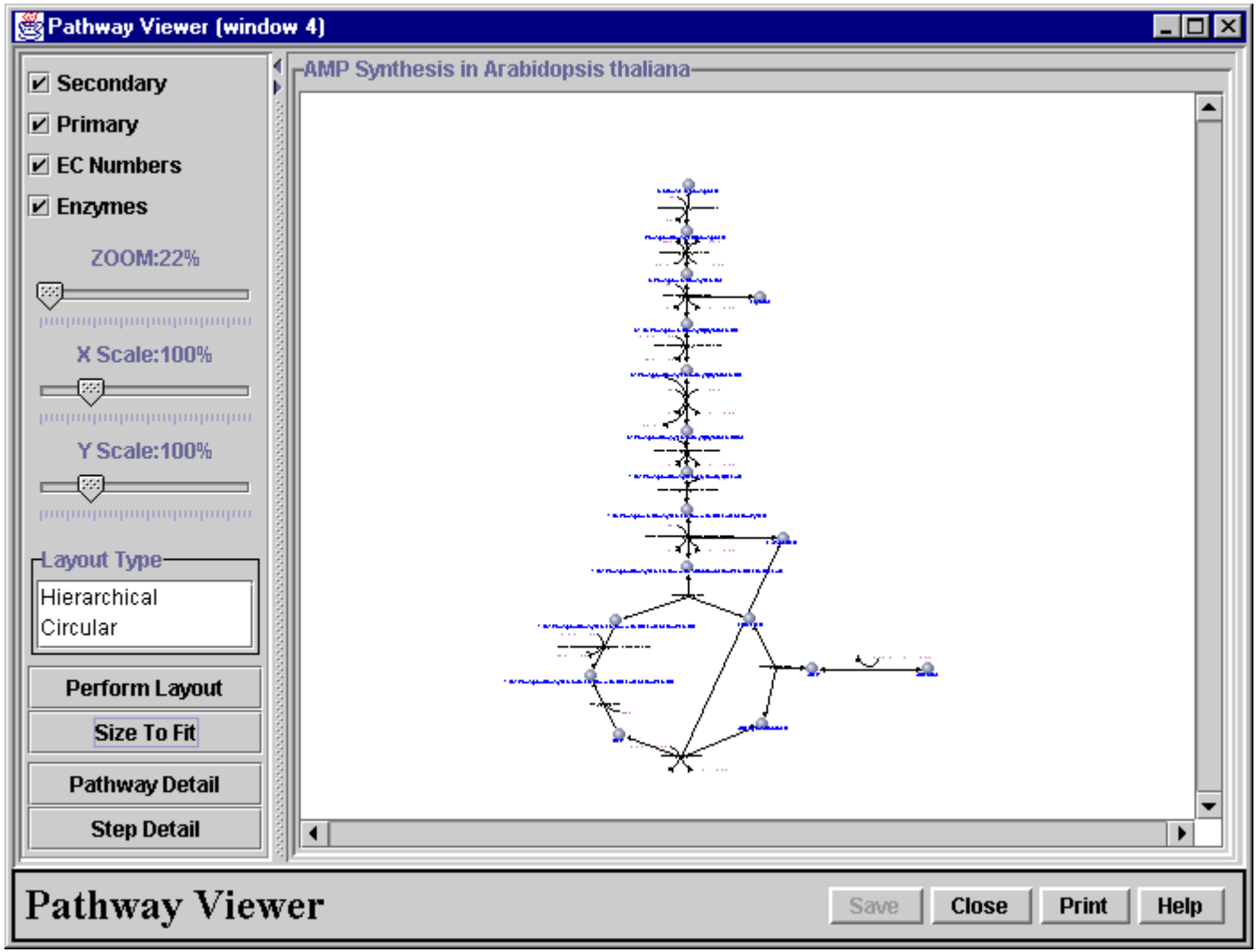
```
HOMO SAPIENS    ------------------------------------------------
MOUSE           AAAP-VVGAVQPVPGMPPMPQAPRIMHHMPGQPPYMPPPGMIPPPGLAPG 183
XENOPUS         ALLPGVPGQMAAMQDMPGMTQAPRMMH-MAGQAPYMHHPGMMPPPGMAPG 178
HOUSE FLY       ----------------------------------------PPKPAPG 137

HOMO SAPIENS    ----------------MAPAQPLSENPPNHILFLTNLPEETNELMLSMLF 34
MOUSE           QIPPGAMPPQQLMPGQMPPAQPLSENPPNHILFLTNLPEETNELMLSMLF 233
XENOPUS         QMPPGGMPHGQLMPGQMAPMQPISENPPNHILFLTNLPEETNELMLSMLF 228
HOUSE FLY       TDEKKDK-KKKPSSAENSNPNAQTEQPPNQILFLTNLPEETNEMMLSMLF 186
                                .  :. :*:***:*************:******

HOMO SAPIENS    NQFPGFKEVRLVPGRHDIAFVEFDNEVQAGAARDALQGFKITQNNAMKIS 84
MOUSE           NQFPGFKEVRLVPGRHDIAFVEFDNEVQAGAARDALQGFKITQNNAMKIS 283
XENOPUS         NQFPGFKEVRLVPGRHDIAFVEFDNEVQAGAARESLQGFKITQSNSMKIS 278
HOUSE FLY       NQFPGFKEVRLVPNRHDIAFVEFTTELQSNAAKEALQGFKITPTHAMKIT 236
                ************.********* .*:*:.**::******** .::***:
```


Horse
Donkey
Pig
Kangaroo
Rabbit
Duck
Pigeon
Chicken
Penguin
Turtle
Dog
Human
Monkey
Snake
Tuna
Moth
Screw-worm fly
Candida
Saccharo-myces
Neurospora
.9
.1
2.9
1.3
2.7
1.7
3
1.4
.6
.8
.2
6.9
1.6
4.6
1.2
1.1
.5
1
1.1
4.9
1.4
3.3
1.1
3.3
.2
16.5
17.2
6.5
9.9
17.4
23.4
5.7
5.4
28.1
15.2
9.6
2.1
Phylogeny


Millions of years
25
20
15
10
5
0
0.36
Pygmy chimp
0.44
0.37
Common chimp
0.36
0.84
Human
0.56
0.72
1.11
Gorilla
1.85
Orangutan
2.30
Gibbon

Microarray Technology




Pathway Viewer (window 4)
Secondary
Primary
EC Numbers
Enzymes
ZOOM:22%
X Scale:100%
Y Scale:100%
Layout Type
Hierarchical
Circular
Perform Layout
Size To Fit
Pathway Detail
Step Detail
AMP Synthesis in Arabidopsis thaliana
Pathway Viewer
Save
Close
Print
Help

GENOMES to LIFE

BIOLOGICAL SOLUTIONS FOR ENERGY CHALLENGES

INNOVATIVE APPROACHES ALONG UNCONVENTIONAL PATHS

CELL

ECOLOGY FOR THE 21ST CENTURY

DNA SEQUENCE DATA FROM GENOME PROJECTS

Genes and other DNA sequences contain instructions on how and when to build proteins

goal IDENTIFY PROTEIN MACHINES

PROTEINS

Proteins perform many of life's most essential functions. To carry out their specific roles, they often work together in the cell as protein machines.

goal CHARACTERIZE GENE REGULATORY NETWORKS

Many protein machines interact through complex, interconnected pathways. Analyzing these dynamic processes will lead to models of life processes.

goal DEVELOP COMPUTATIONAL CAPABILITIES TO UNDERSTAND COMPLEX BIOLOGICAL SYSTEMS

WORKING CELL

COMMUNITY OF CELLS

goal EXPLORE FUNCTION IN MICROBIAL COMMUNITIES

Protect workers and the public

Clean up the environment

Apply knowledge of microbial functional capabilities

Sequester excess carbon

Produce and use energy

URL DOEGenomesToLife.org

10/01

Phage Design

- Bacteria vs. Virus.
- A Phage is a virus that infects bacteria.
- Phages kill a host by reproducing inside them using the hosts machinery.
- Bacteria fight phages by “cutting ‘em up” using Restriction Enzymes.
- Phage Therapy

Restriction Enzymes

- ***Eco*RI** recognizes

  AATT

  and cuts at that site.
- Idea: Modify the phage sequence so that ***Eco*RI** cannot recognize it any more.

Modifying Phages

… CAC TGG TAC TAC CAA TTA CGG CTA …

… CAC TGG TAC TAC CAA TTA CGG CTA …

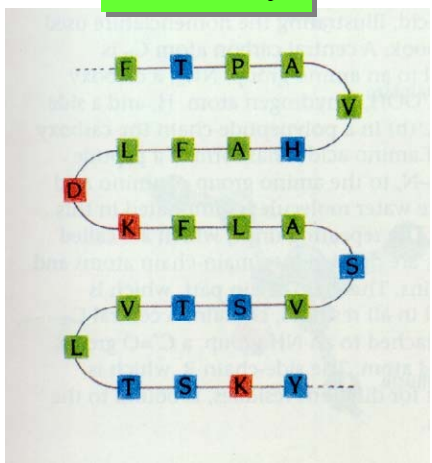
… **H** **W** **Y** **Y** **Q** **L** **R** **L** …

… CAC TGG TAC TAC CAG TTA CGG CTA …

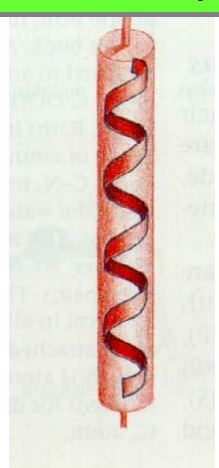
Proteins

- Protein sequences are strings from a 20-letter alphabet.
- Proteins are composed of a sequence of **amino acids**.

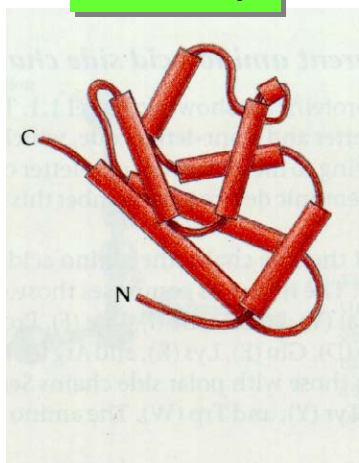
Primary



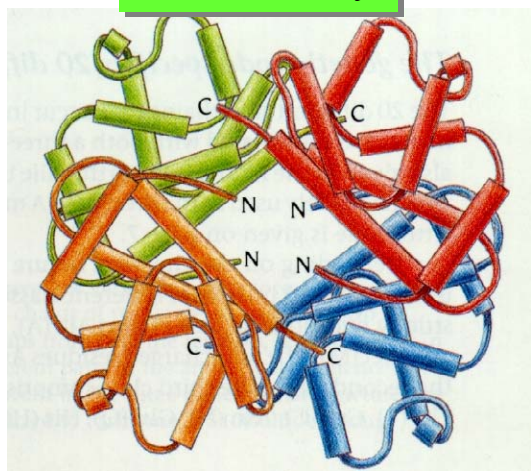
Secondary



Tertiary



Quaternary



Motifs in Protein Sequences

Motifs are combinations of secondary structures in proteins with a specific **structure** and a specific **function**.

Examples: Helix-Turn-Helix, Zinc-finger, Homeobox domain, Hairpin-beta motif, Calcium-binding motif, Beta-alpha-beta motif, Coiled-coil motifs.

Motif Detection Problem

Input: Set, S, of known (aligned) examples of a motif M,
A new protein sequence, P.

Output: Does P have a copy of the motif M?

Example: Zinc Finger Motif

...**Y**K**C**GL**C**ERS**F**VEKSA**L**SR**H**ORV**H**KN...

3 6 19 23

Input: Database, D, of known protein sequences,
A new protein sequence, P.

Output: What interesting patterns from D
are present in P?

Protein Structure Prediction Problem

Input: A given protein sequence, P.

Output: The 3D structure of P.

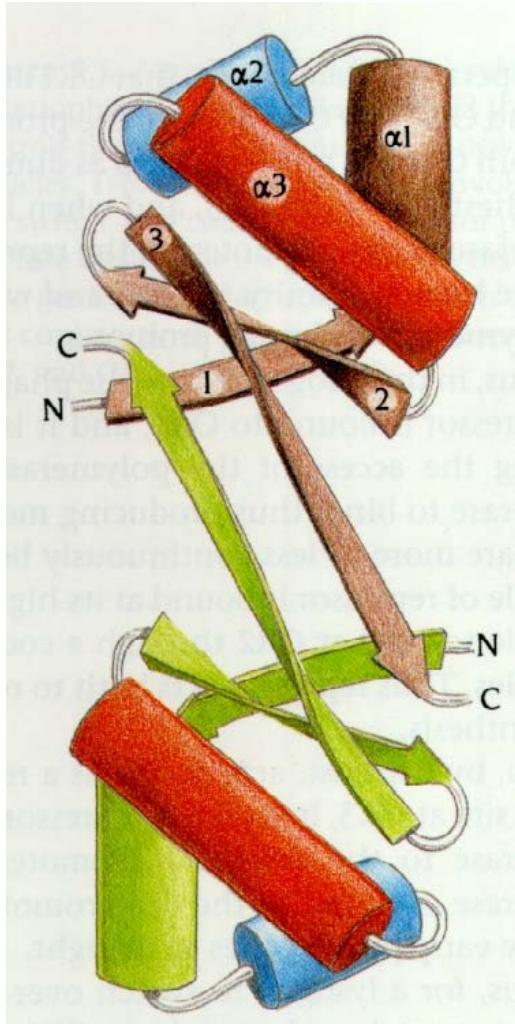
Protein Function Prediction Problem

Input: A given protein sequence, P.

Output: The functional characterization of P.

Helix-Turn-Helix Motifs

- Structure
  - 3-helix complex
  - Length: 22 amino acids
  - Turn angle
- Function
  - Gene regulation by binding to DNA



DNA Binding at HTH Motif

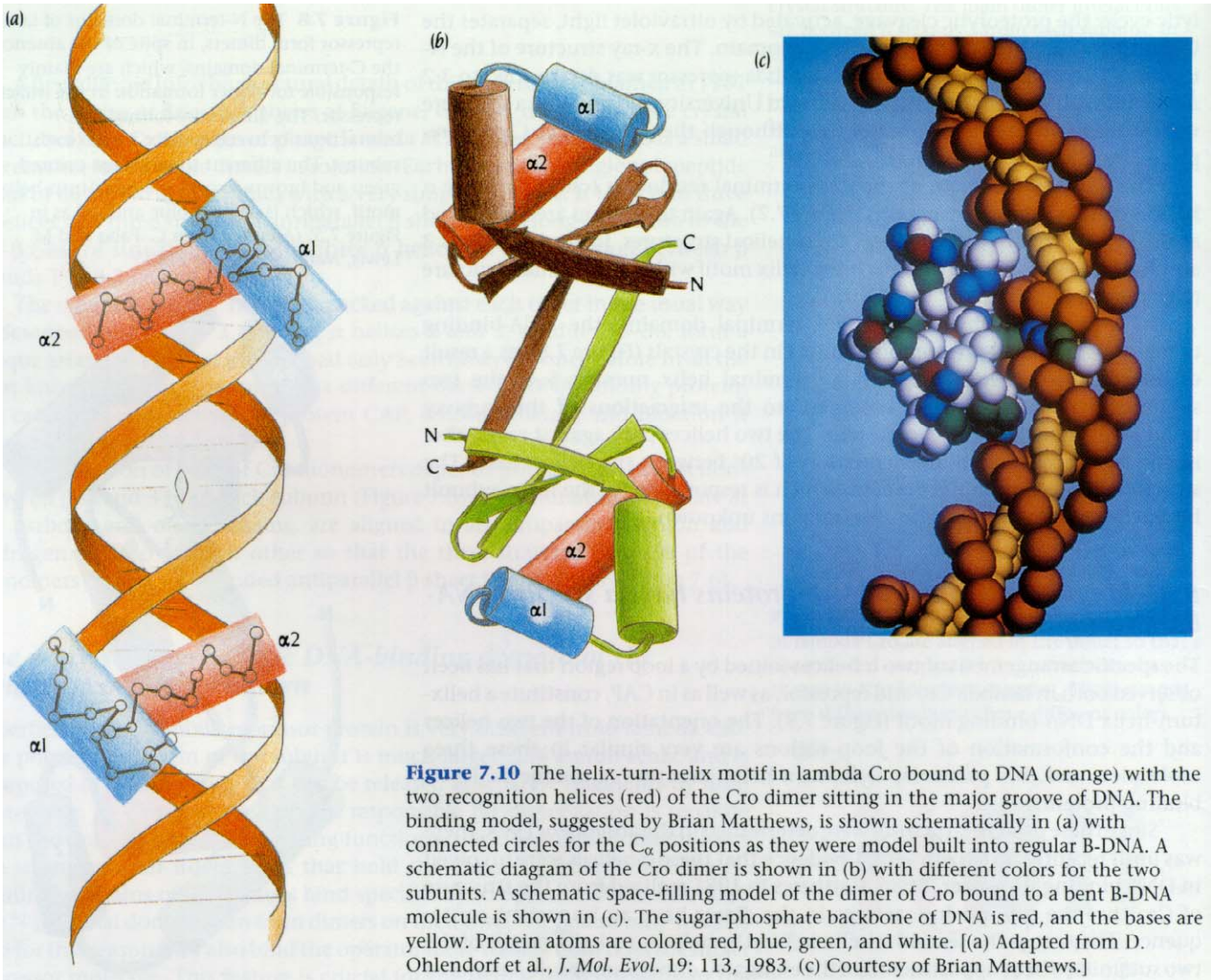


Figure 7.10 The helix-turn-helix motif in lambda Cro bound to DNA (orange) with the two recognition helices (red) of the Cro dimer sitting in the major groove of DNA. The binding model, suggested by Brian Matthews, is shown schematically in (a) with connected circles for the C_α positions as they were model built into regular B-DNA. A schematic diagram of the Cro dimer is shown in (b) with different colors for the two subunits. A schematic space-filling model of the dimer of Cro bound to a bent B-DNA molecule is shown in (c). The sugar-phosphate backbone of DNA is red, and the bases are yellow. Protein atoms are colored red, blue, green, and white. [(a) Adapted from D. Ohlendorf et al., *J. Mol. Evol.* 19: 113, 1983. (c) Courtesy of Brian Matthews.]

HTH Motifs: Examples

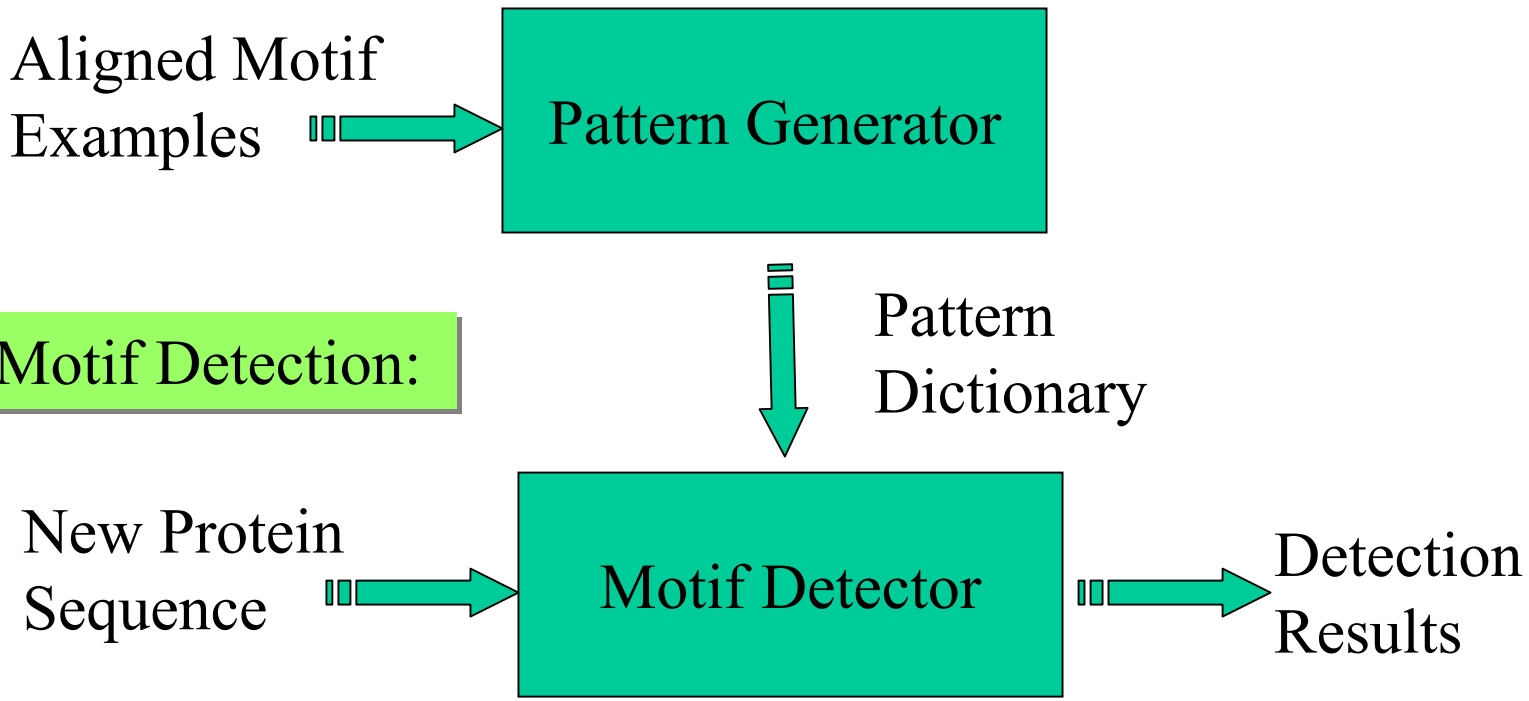
Loc	*Protein*	*Helix 2*									*Turn*				*Helix 3*									
	Name	-1	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
14	**Cro**	F	G	Q	E	K	T	A	K	D	L	G	V	Y	Q	S	A	I	N	K	A	I	H
16	**434 Cro**	M	T	Q	T	E	L	A	T	K	A	G	V	K	Q	Q	S	I	Q	L	I	E	A
11	**P22 Cro**	G	T	Q	R	A	V	A	K	A	L	G	I	S	D	A	A	V	S	Q	W	K	E
31	**Rep**	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

Basis for New Algorithm

- Combinations of residues in specific locations (may not be contiguous) contribute towards stabilizing a structure.
- Some reinforcing combinations are relatively rare.

New Motif Detection Algorithm

Pattern Generation:



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

Experimental Results: **GYM 2.0**

Motif	*Protein Family*	*Number Tested*	*GYM = DE Agree*	*Number Annotated*	*GYM = Annot.*
HTH Motif (22)	Master	88	88 (100 %)	13	13
	Sigma	314	284 + 23 (98 %)	96	82
	Negates	93	86 (92 %)	0	0
	LysR	130	127 (98 %)	95	93
	AraC	68	57 (84 %)	41	34
	Rreg	116	99 (85 %)	57	46
	Total	675	653 + 23 (94 %)	289	255 (88 %)

Experiments

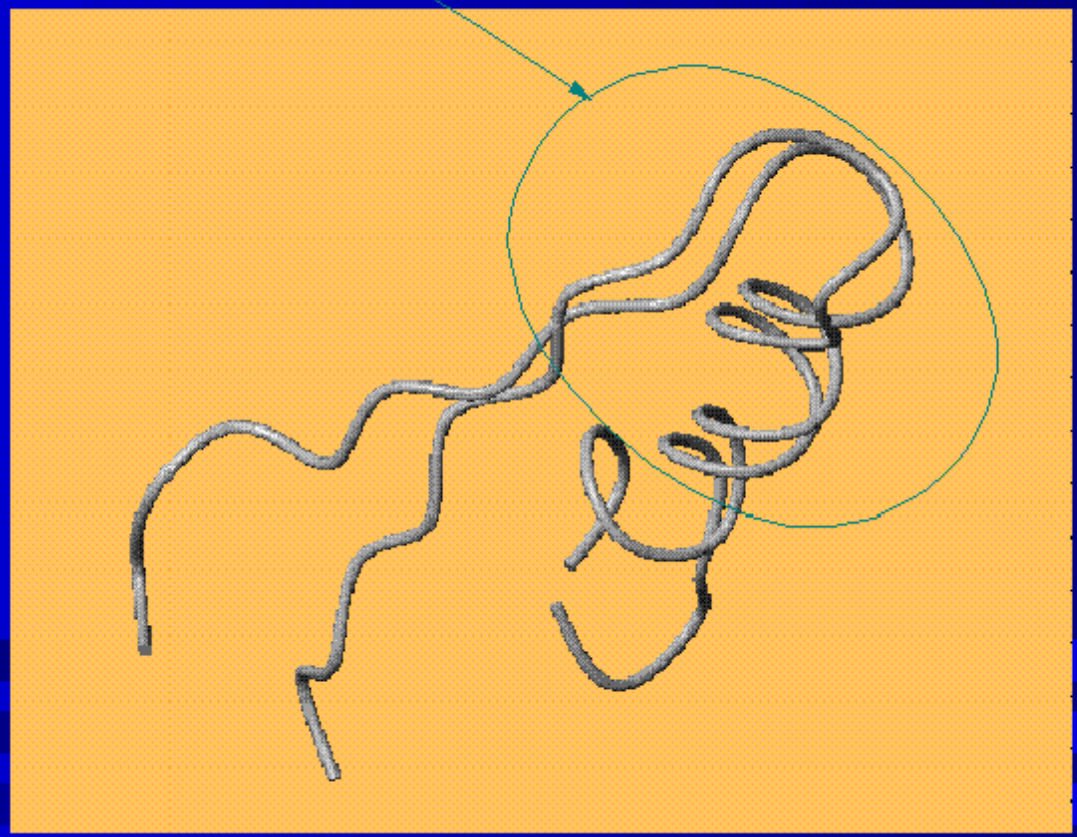
- Basic Implementation (Y. Gao)
- Improved implementation & comprehensive testing (K. Mathee, GN).
- Implementation for homeobox domain detection (X. Wang).
- Statistical methods to determine thresholds (C. Bu).
- Use of substitution matrix (C. Bu).
- Study of patterns causing errors (N. Xu).
- Negative training set (N. Xu).
- NN implementation & testing (J. Liu & X. He).
- HMM implementation & testing (J. Liu & X. He).
- Structlet Assembly & testing (G. Zheng).

Seqlets Describe 3-Dimensional Structure

V..G..G.G.T.L

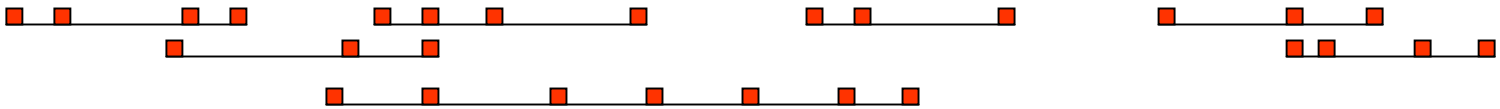
>1ayl
VFFGLSGTGKTTL
>1pox
VCFGSAGPGGTHL

RMS error=
2.192 Angstroms



(Global) Protein Structure Prediction

Gao, 2001



3D-BBAssembly: Assembles structlets with best alignment for overlapping regions. Zheng, 2001

(Global) Protein Structure Prediction

- Find all seqlets from the Seqlet Dictionary present in the given protein sequence.
- List out all structlets from the Structlet Dictionary corresponding to the seqlets in the protein.
- Extend structlet structures using **3D-SLAM**.
- Use branch-and-bound techniques to Assemble structlets to predict global structure. Use consistency of overlap regions to eliminate possible structures.
- Use energy function methods to refine global structure.

Pattern Discovery Applications in Bio-informatics

- Motif Discovery in Proteins
- Single & Composite Descriptors of Protein Families
- Protein Structure Prediction
- Discovery of Tandem Repeats in DNA sequences
- Multiple Sequence Alignment
- Homology Detection; Annotations
- Gene Expression Analysis

Credits

PhD Students:

- **Yuan Gao (2001, IBM T.J. Watson)**
- **Gaolin Zheng, Tom Milledge, Patricia Buendia, Chengyong Yang**

Masters Students:

- **Changsong Bu (Idax)**
- **Xuning Wang (Parke Davis)**
- **Ning Xu (ClonTech)**
- **Gaolin Zheng (FIU)**
- **Peter Dimitrov (Novartis)**
- **Xiao-rui He**
- **Junmin Liu**
- **Meera Krishnan**
- **Hari Tammana (Affymetrix)**
- **Eric Wu**

Collaborators

- **Kalai Mathee**, **Rene Herrera**, **Lydia Kos** (FIU)
- **Isidore Rigoutsos** (IBM T.J.Watson)
- **V. Milenkovic, R. Bookman** (U Miami)
- **Tom Sutter, E. O. George, A. Quas** (U Memphis)
- **M. Li** (U Tennessee)
- **S. Samant** (St. Jude Research Hospital)