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)

...TTCTGCATTCGGTGAAGAGGCAGCTCTAG...
...AAGACGTAAGCCACTTCTCCGCGGAGATC...

Genome
Molecular Biology Primer (Cont’d)

DNA → code for Proteins

DNA → mRNA → Proteins

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

Proteins:
Hemoglobin,
Immunoglobulin,
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):

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

<table>
<thead>
<tr>
<th>1st position (5' end)</th>
<th>2nd position</th>
<th>3rd position (3' end)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>U</strong></td>
<td><strong>C</strong></td>
<td><strong>A</strong></td>
</tr>
<tr>
<td>Leu</td>
<td>Ser</td>
<td>Tyr</td>
</tr>
<tr>
<td>Leu</td>
<td>Ser</td>
<td>Cys</td>
</tr>
<tr>
<td>Leu</td>
<td>Ser</td>
<td>STOP</td>
</tr>
<tr>
<td>Leu</td>
<td>Ser</td>
<td>STOP</td>
</tr>
<tr>
<td>Leu</td>
<td>Pro</td>
<td>His</td>
</tr>
<tr>
<td>Leu</td>
<td>Pro</td>
<td>Arg</td>
</tr>
<tr>
<td>Leu</td>
<td>Pro</td>
<td>His</td>
</tr>
<tr>
<td>Leu</td>
<td>Pro</td>
<td>Arg</td>
</tr>
<tr>
<td>Ile</td>
<td>Thr</td>
<td>Asn</td>
</tr>
<tr>
<td>Ile</td>
<td>Thr</td>
<td>Ser</td>
</tr>
<tr>
<td>Ile</td>
<td>Thr</td>
<td>Asn</td>
</tr>
<tr>
<td>Met</td>
<td>Thr</td>
<td>Lys</td>
</tr>
<tr>
<td>Val</td>
<td>Ala</td>
<td>Asp</td>
</tr>
<tr>
<td>Val</td>
<td>Ala</td>
<td>Gly</td>
</tr>
<tr>
<td>Val</td>
<td>Ala</td>
<td>Gly</td>
</tr>
<tr>
<td>Val</td>
<td>Ala</td>
<td>Gly</td>
</tr>
</tbody>
</table>

**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 GTRVRAMAIYQSQHMTEVVRRCPHHE--RCSDSGLAPPQHLIRVEGNLRVEYLDNRNT 211
  G  +RAM +YK+++H+TEVV+RCP+HE R  +  +APP HLIRVEGN +Y++D T
Sbjct: 128 GAVIRAMPVYKAEHVTEVVKRCNPHELSEFNEGQIAPPSHLIRVEGNSHAQYVEDPIT 187

Query: 212 FRHSVVPYEPPEVGSDCTTIHYNYMCNSSCMGGMNRRPILTIIITLDEDSSGNNLLGRNFE 271
  R SV+VPYEPP+VG++ TT+ YN+MCNSSC+GGMNRRPIL I+TLE G +LGR FE
Sbjct: 188 GRQSVLVPYEPPQVGTEFTTLYNFMCNSSCVGGMNRRPILIIVTLETRDGVLGRRCFE 247
...
## Multiple Sequence Alignment

<table>
<thead>
<tr>
<th></th>
<th>HOMO SAPIENS</th>
<th>MOUSE</th>
<th>XENOPUS</th>
<th>HOUSE FLY</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AAAP-VVGAVQPVPMPMPQAPRIMHHMPQPPYMPPPGMIPPPGLAPG 183</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>ALLPGVPQMAAMQDMPGTMQAPRMMH-MAGQAPYMHPPGMMPPPGMAPG 178</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>--------------</td>
<td>-------</td>
<td>---------</td>
<td>-----------</td>
</tr>
<tr>
<td></td>
<td>HOMO SAPIENS</td>
<td>MOUSE</td>
<td>XENOPUS</td>
<td>HOUSE FLY</td>
</tr>
<tr>
<td></td>
<td>--------------</td>
<td>-------</td>
<td>---------</td>
<td>-----------</td>
</tr>
<tr>
<td></td>
<td>NQFPFGKEVRLVPGHRDHIAFVEFDNEVQAGAARDALQGFKITQNNAMKIS 84</td>
<td>NQFPFGKEVRLVPGHRDHIAFVEFDNEVQAGAARDALQGFKITQNNAMKIS 283</td>
<td>NQFPFGKEVRLVPGHRDHIAFVEFDNEVQAGAARSLQGFKITQNSMKIS 278</td>
<td>NQFPFGKEVRLVPNRHDIAFVEFTTELQSNAAKEALQGFKITPTHAMKIT 236</td>
</tr>
<tr>
<td></td>
<td>--------------</td>
<td>-------</td>
<td>---------</td>
<td>-----------</td>
</tr>
</tbody>
</table>

---
Phylogeny
Microarray Technology

72 hrs

72 hrs + Ganciclovir
GENOMES to LIFE
BIological Solutions for Energy Challenges
Innovative Approaches along Unconventional Paths

DNA Sequence Data from Genome Projects

Goal: Examine Function in Microbial Communities

- Protect workers and the public
- Clean up the environment
- Sequester excess carbon
- Produce and use energy

Goal: Explore Function in Microbial Communities

- Apply knowledge of microbial functional capabilities

Goal: Develop Computational Capabilities to Understand Complex Biological Systems

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

Goal: Identify Protein Machines

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

Goal: 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

URL: DOEGenomesToLife.org
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

- *EcoRI* recognizes
  
  AATT

  and cuts at that site.

- Idea: Modify the phage sequence so that *EcoRI* 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**.
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

```
...YKCGLCERSVEKSAHRLHORVHKN...
```

**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 Cα 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

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

Aligned Motif Examples → Pattern Generator

Motif Detection:

New Protein Sequence → Motif Detector → Detection Results

Pattern Dictionary
# Patterns

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

- **Q1 G9 N20**
- **A5 G9 V10 I15**
# Experimental Results: GYM 2.0

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

VFFGLSGTGKTL

>lpoxl

VCFG SagaPggTfl

RMS error=
2.192 Angstroms
(Global) Protein Structure Prediction

Gao, 2001

3D-BBA\textbf{assembly}: 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)