Computational Biology
From The Perspective Of
A Physical Scientist

Dr. Arthur Dong

PP1@TUM
26 November 2013
Bioinformatics Education

Curriculum

- Math, Physics, Computer Science (Statistics and Programming)
- Life Sciences (Biochemistry and Molecular Biology)
- Bioinformatics (sequence, structure, system)

- Three pillars roughly equal weight; adjustment possible.
- Synergistic collaboration among SEM faculties / schools / universities.

- Appreciate complementary strengths + Acquire complementary ways of thinking.
- Technical competence + Ability to ask the right questions.

Recruitment

- Bioinformatics students from first year on – Marketing!
- Converts from physical sciences:
  - Where else can you do cutting-edge research that matters so early?! 
  - Risk not taking the time to truly understand biology.
- Converts from life sciences:
  - How about finishing an experiment in days rather than months?! 
  - Could be a steep learning curve at the beginning.
- Women – You can excel in this!
How To Read A Paper

Focus: Technical details or the big picture?

**Within the paper:**

- What's the whole point, the *take-home lesson*?
- Why did they do what they did? (historical perspective)
- Any parts problematic and could be improved?
- Expected versus unexpected

**Go beyond the paper:**

- Observation – Question – Hypothesis – Investigation – Application
- What's the next obvious step?
- Can I apply the same ideas/techniques in other areas?

**Turn any question into a project (and possibly a paper)!**
“Classical” Biology (up to 1950s)

- Anatomy – Organs, tissues, cells
- Mendelian Genetics
- Evolution of species

Then came triumph of reductionism...

Somewhere, something went terribly wrong
Where does bioinformatics come into the picture?

Classically:
Protein structure prediction

Genomics:
Sequence search and comparison

Functional genomics and proteomics:
Networks and systems

“Modern” Biology – Molecular Biology of The Cell

- Cell Biology
- Biochemistry
- Molecular Genetics
- Molecular Evolution

The Central Dogma of Molecular Biology

- Replication
  DNA duplicates
- Transcription
  RNA synthesis
- Translation
  Protein synthesis

DNA

RNA

RNA polymerase

Nucleus

Nuclear membrane

Cytoplasm

protein

mRNA

ribosome

mRNA
AGAGCATGTTGGCCTGGTCCTTT
GCTAGGTACTGTAGAGCAGGTGA
GAGAGTGAGGGGGAAGGACTCC
AAATTAGACCAGTTCTTAGCCATG
AAGCAGAGACTCTGAAGCCAGAC
TACCTGGGTCCCAATCTTGGGCTT
GGTATTTCCTCGCTGTGTGACTCT
GGGTAAGTTACTTAACTTCTCTGT
GCCTCAGTTTCTCTCAAGTGTAAAG
TGACGCTTTGAAAAGTGTCCTCCTG
CAAAAAGAAAGGGCTGCTGGGAGG
AGGGGTGTCCCTGTGTGACTA
AGTACAATATGAGGTTTGT

MGLSDGEWQLVLNVWGKVEADIP
GHGQEVLIRLFKGHPETLEKFDKFK
HLKSEDEMKASEDLKKHGATVLTAL
GGILKKKGHHEAEIKPLAQSHATKH
KIPVKYLEFISECIIQVLQSKHPGDF
GADAQGAMNKALELFRKDMASNY
KELGFQG

Genetic Code

AGAGCATGTTGGCCTGGTCCTTT
GCTAGGTACTGTAGAGCAGGTGA
GAGAGTGAGGGGGAAGGACTCC
AAATTAGACCAGTTCTTAGCCATG
AAGCAGAGACTCTGAAGCCAGAC
TACCTGGGTCCCAATCTTGGGCTT
GGTATTTCCTCGCTGTGTGACTCT
GGGTAAGTTACTTAACTTCTCTGT
GCCTCAGTTTCTCTCAAGTGTAAAG
TGACGCTTTGAAAAGTGTCCTCCTG
CAAAAAGAAAGGGCTGCTGGGAGG
AGGGGTGTCCCTGTGTGACTA
AGTACAATATGAGGTTTGT

Genetic Code
Protein Folding and Structure Prediction

Sequence determines structure and structure determines function (roughly!)

Challenge:
  Given target sequence, predict target structure

Homology Modelling:
  Target sequence has a homologous sequence with solved structure
  1. Align the two sequences (crucial step)
  2. Put target sequence onto homologous structure and “massage”
  Need at least 40% homology
What if no close homologous structure?

...QNVERLRLKNHLYSPASFKRLSRLQYLDLHNNFKEIPYILT...

Threading:

- Inverse folding problem
- 3D profile and pairwise contact potential
- Difficulty with multi-domain proteins or those with no clear domain structures
Ab Initio Prediction

Molecular Dynamics

Physics – throw in electric charge, solvent etc. and minimize the energy function

“Logo” Method

Assemble library fragments

- Partial success on small proteins
- In general computationally prohibitive
- How does nature work?

Bradley et al, Science 2005
(Traditional) Tenets of Molecular Biology

One gene, one protein, one function (or disease)

Protein sequence determines structure, structure determines function

- Partial success on small proteins, but dead end?
- How does nature work?

Coiled coils
Beta barrels
Intrinsic disorders
……
Genomics – Producing the “Parts List”
Large-scale sequencing of genomes and the resulting data explosion

**Sequence Comparison:**
- Given a query, find “similar” sequences among tens of millions in databases – fast!
- Align a group of related sequences; identify conserved residues or regions for structure or function prediction.
- Cluster sequences according to different features.

String comparison algorithms and machine learning (regression, clustering, hidden Markov models, neural networks, etc.)

Dynamic programming: optimal alignment

---

**Dynamic programming: optimal alignment**

<table>
<thead>
<tr>
<th>Pair of protein sequences</th>
<th>G G Q L A K E E A L</th>
</tr>
</thead>
<tbody>
<tr>
<td>U</td>
<td>GGQLAKEEAL</td>
</tr>
<tr>
<td>T</td>
<td>EGQPVEVL</td>
</tr>
</tbody>
</table>

**Optimal alignment (no gaps)**

<table>
<thead>
<tr>
<th>U</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>E</td>
</tr>
<tr>
<td>E</td>
<td>Q</td>
</tr>
<tr>
<td>Q</td>
<td>P</td>
</tr>
<tr>
<td>P</td>
<td>V</td>
</tr>
<tr>
<td>V</td>
<td>E</td>
</tr>
<tr>
<td>E</td>
<td>V</td>
</tr>
<tr>
<td>V</td>
<td>L</td>
</tr>
</tbody>
</table>

**Optimal alignment (with gaps)**

<table>
<thead>
<tr>
<th>U</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>E</td>
</tr>
<tr>
<td>E</td>
<td>Q</td>
</tr>
<tr>
<td>Q</td>
<td>P</td>
</tr>
<tr>
<td>P</td>
<td>V</td>
</tr>
<tr>
<td>V</td>
<td>E</td>
</tr>
<tr>
<td>E</td>
<td>V</td>
</tr>
<tr>
<td>V</td>
<td>L</td>
</tr>
</tbody>
</table>

**String Comparison:**
- Given a query, find “similar” sequences among tens of millions in databases – fast!
- Align a group of related sequences; identify conserved residues or regions for structure or function prediction.
- Cluster sequences according to different features.

**String Comparison Algorithms:**
- Dynamic programming
- Local alignment
- Global alignment
- Smith-Waterman algorithm

**Dynamic Programming:**

\[
SW = \sum_{k=1}^{L_{ali}} M_{ij}(U_k, T_k) = Go \cdot N_{gap} - Ge \cdot (L_{gap} - N_{gap})
\]
Functional Genomics and Proteomics
Understanding How Parts Work Individually and Together

- Genome-wide mRNA expression profiling; synthetic lethal screening
- Proteome-wide Yeast-2-hybrid screening and co-AP/MS
Protein-Protein Interactions: Stable Complexes
PPI network is not static!

- Different time
- Different space
- Different conditions
Beyond Physical Interactions

Genetic Landscape Of A Cell
From Chimpanzees to Neandertals to ...
Humans

- 1000 genomes
- Common vs rare, disease-causing SNPs
- Interconnections among diseases
a Human Disease Network

b Disease Gene Network
Cancer Classification

Set Marker (Leukemia)

Network Marker (Breast Cancer)
Looking Ahead

Classical Biology

Molecular Biology

Systems Biology

Whole

Parts

Whole

Understand how life works
Mechanism (and hopefully) treatment for cancer and other diseases
Synthetic biology, new materials, new energies

Genome

Computational Systems Biology

Phenome

What questions to ask?
What stories to tell?