Arts and Ideas

Where DNA and CPU Meet
a presentation on how computer science and biology inform each other

Dr. Ray Klump
Chair, Mathematics & Computer Science
What is the human genome?
DNA is the **blueprint of life.**
Sequencing involves mapping out the order of molecules within DNA.
DNA is a molecule consisting of a long chain of nucleotides labeled A, G, C, and T.
Each nucleotide has a carbon ring as well as another ring of carbon, nitrogen, and oxygen called the base.
DNA chains are made by connecting these molecules together at their phosphates.
Double-stranded DNA forms when two single strands line up at their bases.
The strands line up predictably:

A $\leftrightarrow T$

G $\leftrightarrow C$
The blue middle are the rungs of **bases** that make the double helix.
The “pickiness” of the nucleotides helps us build DNA from existing strands through denaturation, annealing, and replication.
DNA sequencing reactions are similar.
For sequencing, however, the reactions are run in the presence of dideoxynucleotides. These terminate the chains on specific bases.
If dideoxy-C is used ...

If all four dideoxy’s are used

```
GCGATTCGTCGCTACGCTACGCT
GCGATTCGTCGCTACGCTACGCT
GCGATTCGTCGCTACGCTACGCT
GCGATTCGTCGCTACGCTACGCT
GCGATTCGTCGCTACGCTACGCT
GCGATTCGTCGCTACGCTACGCT
GCGATTCGTCGCTACGCTACGCT
GCGATTCGTCGCTACGCTACGCT
```

```
GCGATTCGTCGCTACGCTACGCT
GCGATTCGTCGCTACGCTACGCT
GCGATTCGTCGCTACGCTACGCT
GCGATTCGTCGCTACGCTACGCT
GCGATTCGTCGCTACGCTACGCT
GCGATTCGTCGCTACGCTACGCT
GCGATTCGTCGCTACGCTACGCT
GCGATTCGTCGCTACGCTACGCT
```

```
GCGATTCGTCGCTACGCTACGCT
GCGATTCGTCGCTACGCTACGCT
GCGATTCGTCGCTACGCTACGCT
GCGATTCGTCGCTACGCTACGCT
GCGATTCGTCGCTACGCTACGCT
GCGATTCGTCGCTACGCTACGCT
GCGATTCGTCGCTACGCTACGCT
GCGATTCGTCGCTACGCTACGCT
```
Electrophoresis automates this
Scaling this process to the size of the human genome requires sophisticated computer algorithms.

3 billion bases!
Today’s **computer technology** has made interpreting sequence data possible.
The **US Human Genome Project** uses huge libraries of large human DNA strands to sequence the genome.
The **US Human Genome Project** took 13 years and $3 billion.
The project identified all of the nearly 25,000 genes in human DNA, and it determined the sequences of the 3 billion base pairs that comprise it.
It is estimated that the Human Genome Project returned $140 for every dollar invested.
Today, this same sequencing can be done by a stand-alone laboratory in one day for several thousand dollars.
The $1,000 Genome, and the New Problem of Having Too Much Information

The next sequence is even cheaper

By Jennifer Abbasi  Posted 02.27.2012 at 11:19 am  ▼  5 Comments

Packed Chip The Ion Proton I can sequence much of a human genome for just $1,000. Sequencing will become even cheaper Courtesy Life Technologies
“This technology promises to deliver a complete human genome in 15 minutes.”
Moore’s Law has been good news for this work. A new version will make two-hour, full-genome sequencing possible.

3.2 GB worth of data per person.
Research consortia are using advances in **DNA sequencing** technology to **unlock the mysteries** surrounding diseases and disorders.
Understanding each individual’s DNA sequence carries the promise of personalized medicine.
Example:

Mendelian Disorders
Example:
Cancer Genetics
and the
Cancer Genome Atlas
Example:

St. Jude Children’s Hospital
When *genome sequencing* begins reaching millions of patients, it will help address the most common problems in medicine.
Of course, **how do you store** all that data and retrieve it quickly?
Companies have been created that host genomes on the cloud for scientists and doctors to access.
Doctors will need to be trained to apply genomic information to standard medical practice.
But there’s more!
In addition to genome sequencing, **Computer Science** is being used for a lot of other **biological research** today.
Data mining has been used to determine dangerous drug interactions.
The FDA maintains a database called AERS that lists 4 million negative reactions since 1969, but this data alone doesn’t capture the full complexity of drug interaction.
Other data sources include international side effects databases, social network data, warning labels, electronic medical records, and the drugs’ biological targets.
Data mining alone can’t prove that particular drugs cause particular side effects, but it can provide clues.
Predicting side effects can be done even **before the drug hits the market.**
The Similarity Ensemble Approach (SEA) looks for similarities in the targeted proteins.
For example, a recent study looked at 656 drugs to detect similarities with molecules that bind with 73 different proteins associated with side effects.
They discovered nearly 1,200 new interactions.
Software has been written to model the spread of disease.
Computer Science has also been used extensively to simulate cells, organs, and organisms.
The first comprehensive computational model of a living organism is of Mycoplasma genitalium.
Other projects have concentrated on modeling a specific organ.
President Obama pitches $100 million investment in human brain research

The President first mentioned his plan to invest in brain research in his State of the Union address. He wants the research to involve private institutions as well as government agencies.

President Obama said the so-called BRAIN Initiative may eventually help find cures for disorders like Alzheimer’s, epilepsy and traumatic injuries.
Such models, when combined with genome information, could allow doctors to prescribe the best treatment based on an individual’s personal genome and history.
This is the idea of truly personalized medicine!
Biology is also transforming Computer Science.
DNA computers will be capable of storing billions of times more data than your personal computer.
DNA logic gates have been created to make more general-purpose computers.
The Traveling Salesman Problem
DNA computing was used by Leonard Adleman in 1994 to solve this.

1. Strands of DNA represent the seven cities. In genes, genetic coding is represented by the letters A, T, C and G. Some sequence of these four letters represented each city and possible flight path.

2. These molecules are then mixed in a test tube, with some of these DNA strands sticking together. A chain of these strands represents a possible answer.

3. Within a few seconds, all of the possible combinations of DNA strands, which represent answers, are created in the test tube.

4. Adleman eliminates the wrong molecules through chemical reactions, which leaves behind only the flight paths that connect all seven cities.
Why DNA computing?
Cheaper, smaller, more environmentally friendly computers.
Conclusion

Computer Science and Biology are transforming each other in exciting ways, much of which is related to DNA.