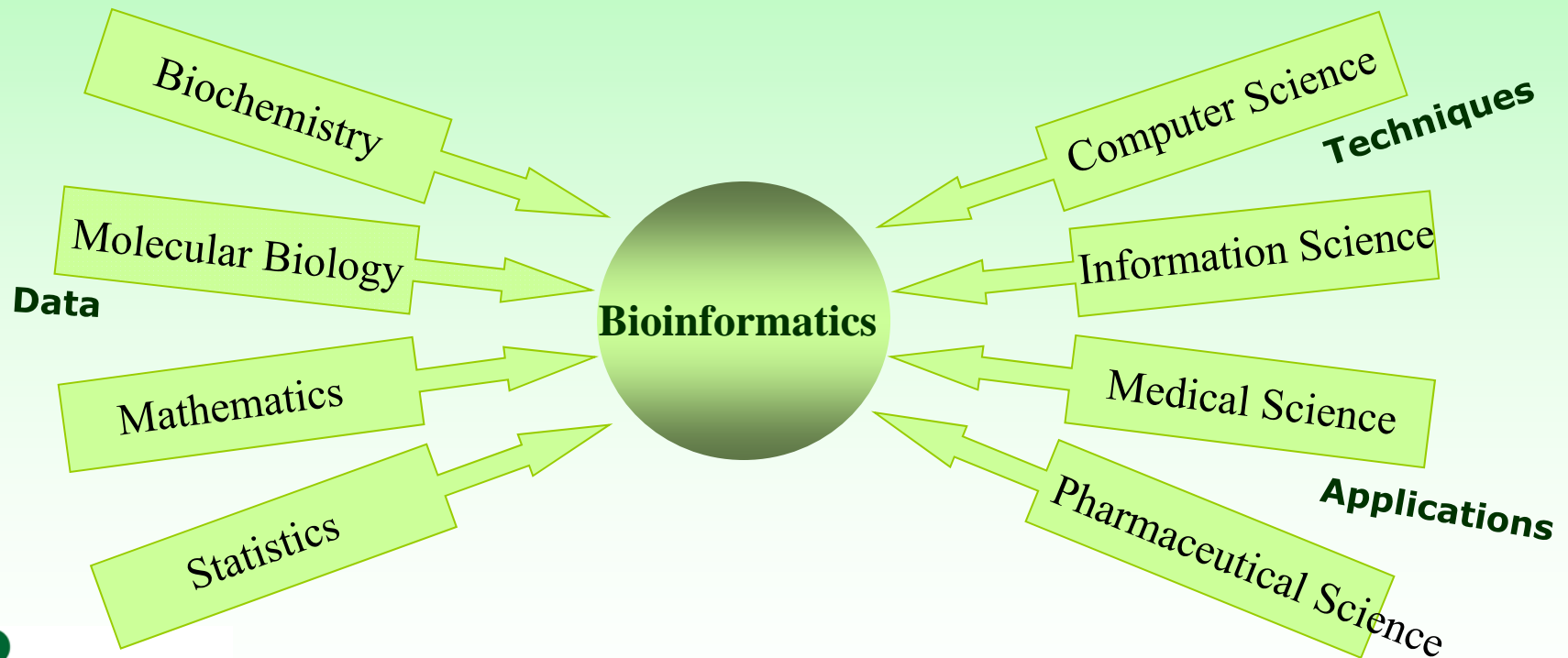


BIOINFORMATICS

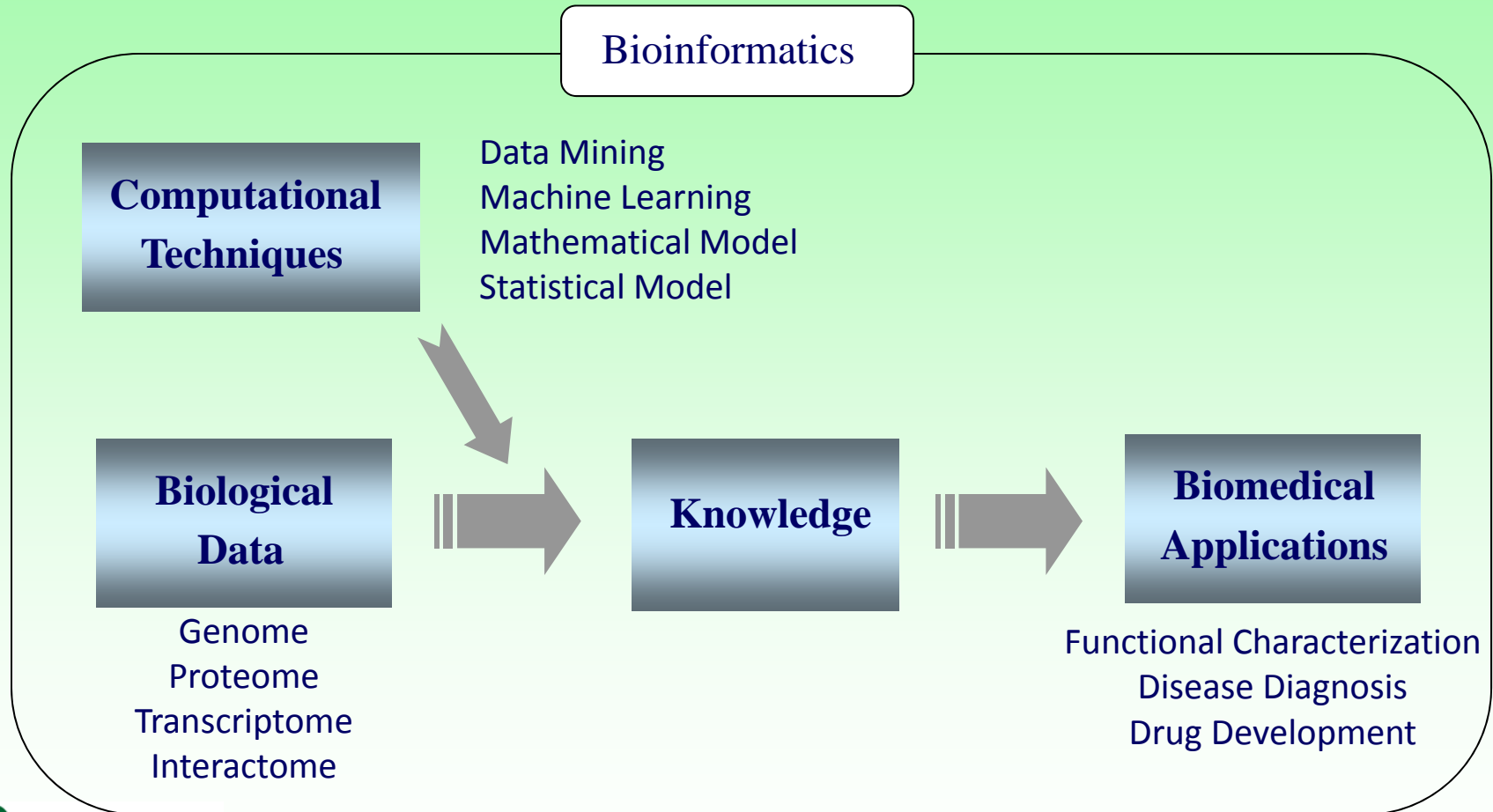
Young-Rae Cho, Ph.D.
Assistant Professor
Department of Computer Science
Baylor University

What is Bioinformatics?

- Interdisciplinary research area to *manage* and *analyze* biological data using computational techniques



What is Bioinformatics?



Bioinformatics

**Computational
Techniques**

Data Mining
Machine Learning
Mathematical Model
Statistical Model

**Biological
Data**

Genome
Proteome
Transcriptome
Interactome

Knowledge

**Biomedical
Applications**

Functional Characterization
Disease Diagnosis
Drug Development

Computational Biology

Stage 1. Sequence Analysis

```
Residue: 16024 (Sequenced Strand) Sequence Length: 1122
16024 TTCTTTCATG GGGAGGAGAG TTTGGGTACC ACCCAAGTAT TGACTCACCC
16074 ATCAACAAAC CCTATGTATT TCGTACATTG CTGCCAGCCA CCAATGAATAT
16124 TGTACGGTAC CATAAATACT TGACCACTTG TAGTACATAA AAACCCAAATC
16174 CACATCAAAA CCCCCCTCCC ATGC TTACAA GCAAGTACAG CAATCAACCC
16224 TCAACTATCA CACATCAACT GCACTCCAAA AGCCACCCCT CCCCCATAG
16274 GATACCAACA AACCTACCCA CCTTAAAGAG TACATAGTAC ATAAAGCAT
16324 TTACCGTACA TAGEACATTA CAGTCAATC CCLTTCCTGC CCATGGATG
16374 ACCCCCTCA GATAGGGGTC CCTTGACCAC CATECTCCTG GAAATCAATA
16424 TCCCGCACAA GAGTGTACT CTCTCCGTC CGGCCCCATA ACATTGGGG
16474 GTAGCTAAAG TGAATGTAT CCGACATCTG GTTCTACTT CAGGGTCATA
16524 AAGGCTAAT AGCCCAACAG TTCGCCTTAA TAAAGCATC AGGATGGATC
16574 ACAGGCTAT CACCCATTA ACCCACTACG GGAGCTCTCC ATGCATTGG
16624 TATTTTCGTC TGGGGGTAT GCACTGGATA GCACTGGAG AGCTGGAGC
16674 CGAGCACCC TATGTCGAG TATCTGCTT TGACTCTCC CTCATCCTAT
16724 TATTATCGC ACCTACGTT AATAATACAG GCAACATAC TTAATAAGT
16774 GTGTAATTA ATTAATGCTT GTAGGACATA ATAAATACAA TTGAATGCTT
16824 GCAACAGCAC TTCCACACA GACATCATAA CAATAAATTT CCAACAAAC
16874 CCCCCCTCCC CGCTTCGAGC CACAGCACTT AAACACATCT CTGCCAACC
16924 CCAAAAACAA AGAACCTTAA CACCAGCTA ACCAGATTG AAATTTTATC
16974 TTTTGGGTT ATGCATTTT AACAGTACC CCCCACATA CACATTTT
17024 TCCCCCCTA CTCCCACTT ACTAATCTCA TCAATAACAC CCCCCCCAT
17074 CCTACCCAGC ACACACACAC CGCTGTAAAC CCAATACCCC GAACCAACCA
17124 AACCCCAAG ACACCCCCCA CA
```

- Gene sequencing
- Sequence alignment
- Homolog search
- Motif finding

Computational Biology

Stage 1. Sequence Analysis

Residue: 16024	(Sequenced Strand)	Sequence Length: 1122
16024	TTC TTT CATG GGG AAG CAGA TTG GGT ACC ACC CAAGTAT TGA CTC ACC	
16074	ATC AAC AAC CC CTA TGT ATT TCG TAC ATTA CTG CAG CCA CCA TGA AT	
16124	TGT GGG TAC CATAAAT ACT TGA CAC CTG TAG TAC ATAA AAA CCA ATC	
16174	CAC AT AAAA CCCCC TCCC ATG TTA CAA GCA GTAC AG CAAT CAAC CC	
16224	TCA ATAT CA CAC ATCA CT GAC TCC AA AGC ACC CT CAC CCA TGG	
16274	GAT ACC A CA AAC CTAC CA CCT TAA CAG TAC ATAG TAC ATA AAG CAT	
16324	TTAC CGTACA TAG C AATTA CAG TCA ATC CTT TCC GTC CCA TGG ATG	
16374	ACCC CCA CA GAT AGGG TCC CTT GAC CAC CATE CCT CGT GAA TCA ATA	
16424	TCCC GCA CA GAT GCT ACT CTC TCC GTC CCG CCA CA ACAT TGG GG	
16474	GTAG CTA AAG TGA ACTGT AT CCG ACCT CTG GTT CCT ACT CAG GGT CATA	
16524	AAG CCT AAT AGCC CAC AG TCCC CTAA TAA GAC ATC CAG GGT GAT C	
5	ACAG GTCTAT CACCC TATTA ACCACT CAG GAG CTCTCC ATG CAT TGG	
85	TAT TTT CGT TGGGG GTAT GCA CGG ATA GCA TCC GAG ACG TGG AGC	
105	CGAG ACC TA TGT CG AG TAT TGT CTT TGA TCC TCC CTA CCT AT	
125	TAT TAT CGC ACC TAC GTT AAT ATAC AG GCA CATA C TAC TAA GT	
205	GTG TTA ATTA ATTA ATG CTT TAG GAC ATA ATA TAA CAA TTG AAT GCT	
225	GCA AGC CAC TTCC CACA GAC ATATA CA AAA ATTT CCA CCA ACC	
305	CCCC CCCC CCG TTT CCG CAC AGC ATT AAAC AAT CT CCG CCA ACC	
385	CCAAA ACAA AGAAC CTA CACC AGCTA ACCAG TTC AAATTT ATC	
465	TTTT GGGT ATG CACTT AAC AGT ACC CCA CATA CAC ATTTT	
545	TCCC CCA C TCCC TACT ACTA ATCTA TCA ATACA C CCCC GCC AT	
625	CCT ACC CAG CAC ACAC AC CGT GCT AAC CCA TAC CCG GAC CA CCA	
705	AACCC CAA G ACAC CCA CA	



Stage 2. Structure Analysis



- Gene sequencing
- Sequence alignment
- Homolog search
- Motif finding
- Protein folding
- Homolog search
- Protein docking
- Function prediction

Human Genome Project

- Goal
 - Identification of complete human genome
 - Mapping the genes from a functional standpoint
- History
 - Began in 1990 by NIH
 - An initial draft was released in 2001
- Participation
 - International Consortium was founded by hundreds of universities and research centers from US, UK, Germany, France, Japan, Canada, China, New Zealand, India, and so on.

Computational Biology

Stage 1. Sequence Analysis

Residue: 16024	(Sequenced Strand)	Sequence Length: 1122
16024	TTC TTT CATG GGG AAG CAGA TTGGGTACC ACCCAAGTAT TGACTACCC	
16074	ATCAACAAACC GCTATGTATT TCGTACATT TGGCAGCCA CCAATGAATAT	
16124	TGTACGGTAC CATAAAATCT TGACCACTG TAGTACATAA AAACCCAAATC	
16174	CACATCAAAA CCCCCCTCCC ATGC TTACAA GCAAGTACAG CAATCAACCC	
16224	TCAACTATCA CACATCACTT GCACTCCAA AGCCACCCCT CACCCACTAG	
16274	GATACCAACA AACCTACCCA CCTTAAAGG TACATAGTAC ATAAAGCAT	
16324	TTACCGTACA TAGCAATTA CAGTCAATC CCLTTCCTGC CCATGGATG	
16374	ACCCCECTCA GATAGGGGTC CCTTGACCAC CATECTCCGT GAATCAATA	
16424	TCCCGCACAA GATGCTACT CTCTCGCTC GGGCCCATAC ACAC TTGGGG	
16474	GTAGCTAAAG TGAATGTAT CCGACATCTG GTTCTACTT CAGGGTCATA	
16524	AAGCCTAAT AGCCCAACAG TTCCECTTAA TAAAGCATC AGGATGGATC	
5	ACAGGCTAT CACCCATTA ACCACTCAGC GGAGCTCTCC ATGCATTGG	
85	TATTTTCGT TGGGGGTAT GCACTGATA GCAATCGGAG ACGCTGGAGC	
105	CGAGCACCC TAGTCCGAG TATCTGCTT TGA TCCCTGC CTCATCCTAT	
125	TATTTATGC ACCTACGTT AATATTACAG GCGAACATC TTACTAAGT	
205	GTGTTAATTA ATTAATGCTT TAGGACATA ATAATAMCAA TTGAATGCTT	
225	GCAAGCCAC TTTCACACA GACATATTA CAATAATTT CCAACAAACC	
305	CCCCCTCCCC CGCTTCGAG CACAGCATTT AAGACATCTT CTGCCAACC	
385	CCAAAACAAA AGAACCTTAA CACCAGCTA ACCAGATT CAAATTTATC	
465	TTTTGGGGT ATGCATTTT AACAGTACC CCCCACATA CACATTTAT	
545	TCCCCCCTCA CTCCCACTC ACTAATCTCA TCAATACAA CCCCCCCAT	
625	CCTACCCAGC ACACACACAC CGCTGCTAAC CCAATACCCC GAACCAACCA	
705	AACCCCAAG ACACCCCCCA CA	



Stage 2. Structure Analysis



- Gene sequencing
- Sequence alignment
- Homolog search
- Motif finding
- Protein folding
- Homolog search
- Protein docking
- Function prediction

Functional Genomics

Stage 1. Sequence Analysis

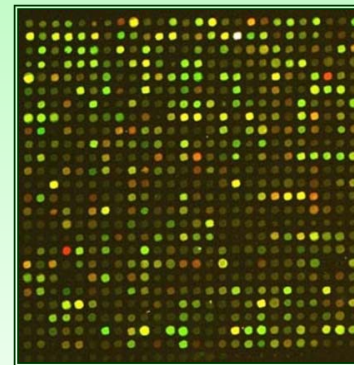
```
Residue: 16024 (Sequenced Strand) Sequence Length: 1122
16024 TTCTTTCATG GGGAGCAGA TTGGGTACC ACCCAAGTAT TGACTACCC
16074 ATCAACACC CTAATGTATT TCGTACATTA CTGCCAGCA CCAATGAATAT
16124 TGTACGGTAC CATAAATACT TGACCACTG TAGTACATAA AAACCCAAATC
16174 CACATCAAAA CCCCCCTCCC ATGC TTACAA GCAAGTACAG CAATCAACCC
16224 TCAATATCA CACATCACTT GCAGTCCAA AGCCACCCCT CACCCACTAG
16274 GATTACCAACA AACCTACCCA CCTTAAAGAG TACATAGTAC ATAAAGCAT
16324 TTACCGTACA TAGEACATTA CAGTCAATC CATTCTCGTC CCTATGGATG
16374 ACCCCCTCA GATAGGGGTC CCTTGACCAC CATECCCTGT GAAATCATA
16424 TCCCGCAAA GATGCTACT CTCTCGTC CCGGCCATA ACATTGGGG
16474 GTAGCTAAG TGAATGTAT CCGACCTTG GTTCTACTT CAGGGTATA
16524 AAGGCTAAT AGCCACAGC TTCCTTAA ATAAGACATC ACGATGGATC
16574 AAGGCTAT CACCCATTA ACCACTCAG GGAAGCTCC ATGCATTGG
16624 TATTTCTG TGGGGGAT CACCGGATA GCAATCGAG ACGTGGAGC
16674 CGAGACCC TATGTCGAG TATCTGCTT TGAATCCGC CTCATCCTAT
16724 TATTTCTG ACCTAGTTC AATATTACAG GCGAACATC TTACTAAGT
16774 GTGTTAATTA ATTAAATCTT TAGGACATA ATAATAMAA TTGAATGCT
16824 GCACAGCAC TTCCACACA GACATATTA CAATAATTT CCAACAAAC
16874 CCCCCCTCCC CGCTTTCGGC CACAGCATTT ABAACATCT CTGCCAACC
16924 CCAAAAACAA AGAACCTTAA CACCAGCTA ACCAGATTC AAATTTATC
16974 TTTTGGGCT ATGCATTTT AACAGTACC CCCCACATA CACATTTAT
17024 TCCCCCCA CTCCCACT ACTAATCTA TCAATAAC CCCCACCAT
17074 CCTACCCAGC ACACACAC CCGTCTAAC CCTATACCC GAACCAACA
17124 AACCCCAAG ACACCCCCA CA
```



Stage 2. Structure Analysis



Stage 3. Genome Analysis



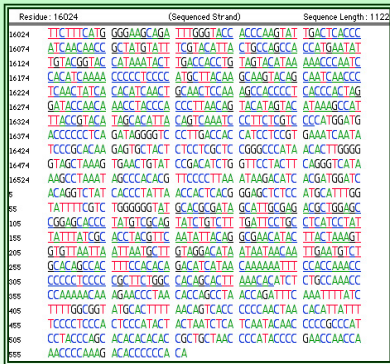
- Gene sequencing
- Sequence alignment
- Homolog search
- Motif finding

- Protein folding
- Homolog search
- Protein docking
- Function prediction

- Function prediction
- Gene clustering
- Disease classification

Systems Biology

Stage 1. Sequence Analysis



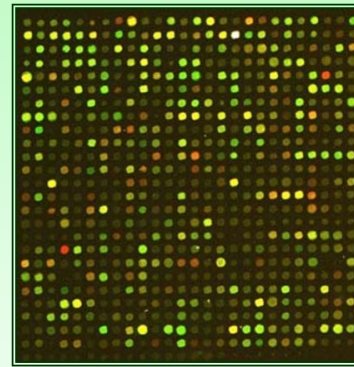
- Gene sequencing
- Sequence alignment
- Homolog search
- Motif finding

Stage 2. Structure Analysis



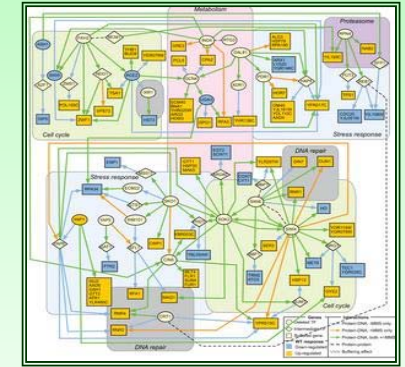
- Protein folding
- Homolog search
- Protein docking
- Function prediction

Stage 3. Genome Analysis



- Function prediction
- Gene clustering
- Disease classification

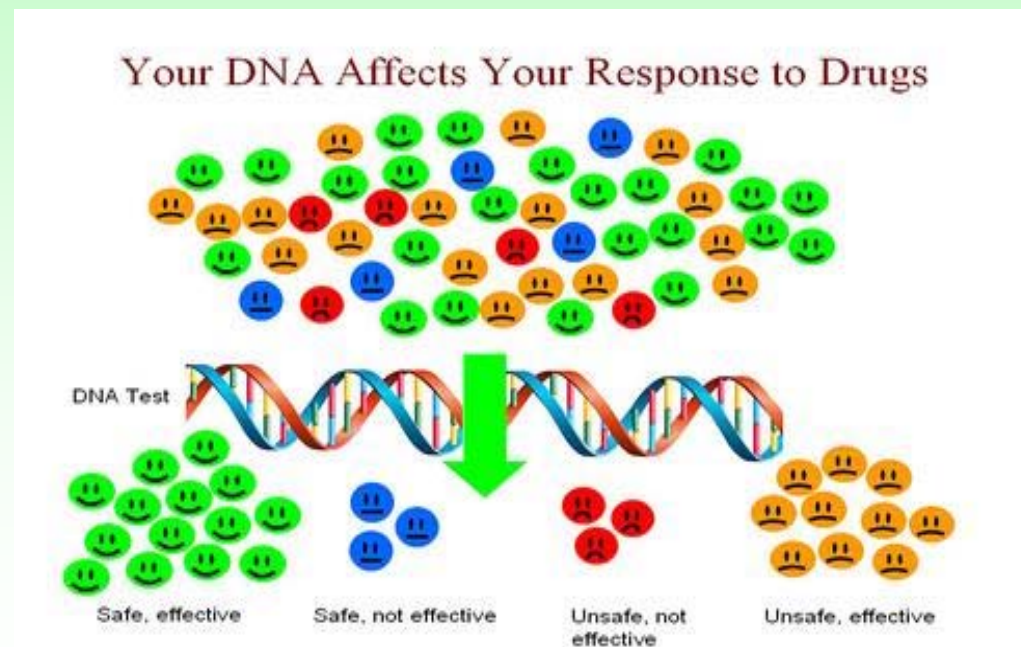
Stage 4. Network Analysis



- Network modeling
- Function prediction
- Pathway identification
- Module detection

Personalized Medicine

- Customized medical treatment
 - examination of genetic variation
 - understanding of the roles of genes in disease



Why Bioinformatics at Baylor?

▪ **Bioinformatics Program in Baylor**

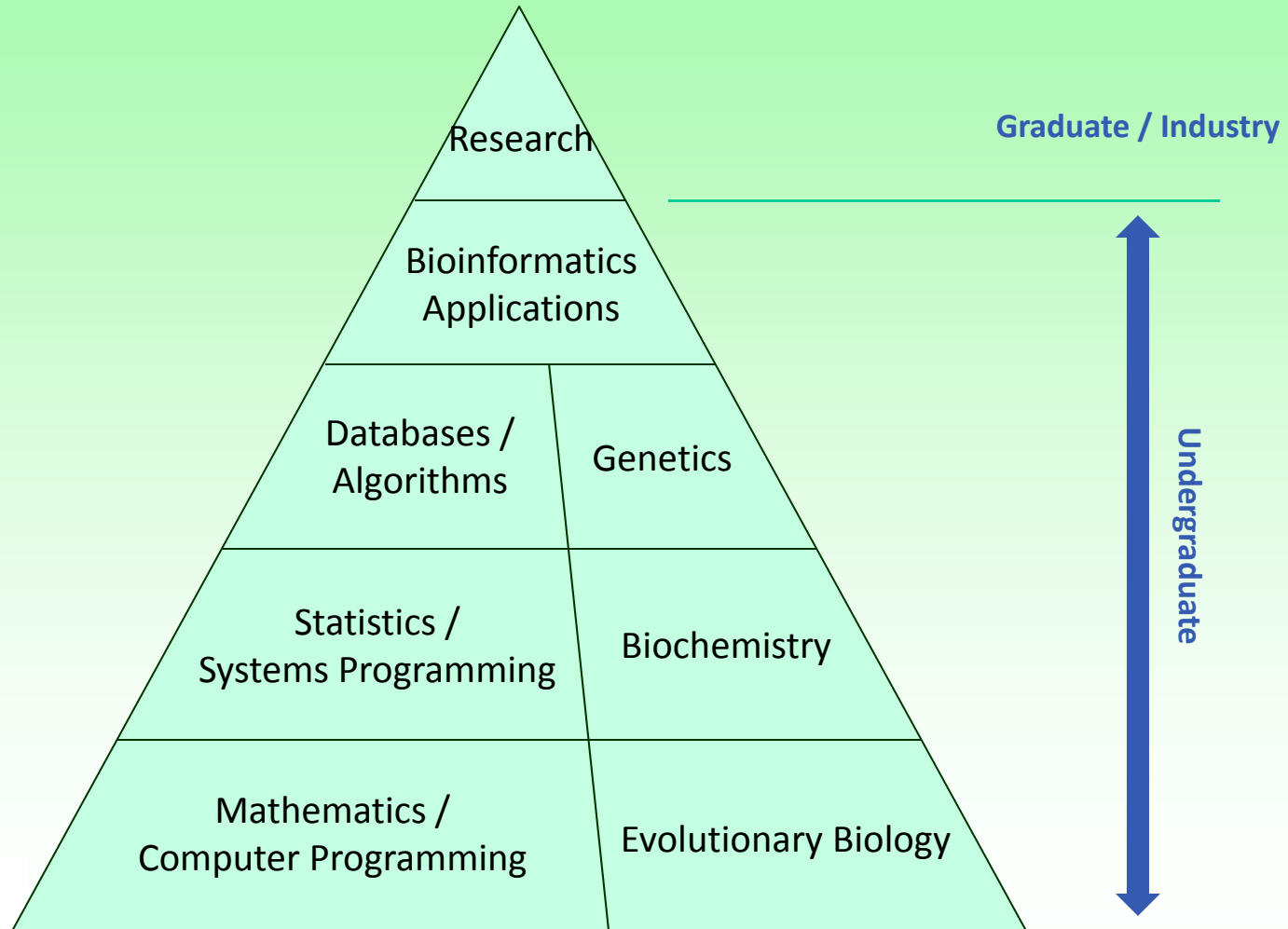
- Bachelor of Science in Informatics (BSI), majoring in Bioinformatics
- First undergraduate program nationally (1999). Continues to occupy a very unique niche.
- There is an increasing need to train individuals at the intersection of computer science, biology (genetics), biochemistry, mathematics, and statistics.
- Two faculty members, Dr. Erich Baker and Dr. Young-Rae Cho
- Approximately 40-50 majors.

Why Bioinformatics at Baylor?

▪ Requirements

- Computer Science: 32 credit hours
- Biology: 20 credit hours
- Chemistry: 17 credit hours
- Bioinformatics: 12 credit hours
 - Genomics & Bioinformatics
 - Introduction to Computational Biology
 - Capstone Bioinformatics Design Project
- General Electives: 0 credit hours

Why Bioinformatics at Baylor?



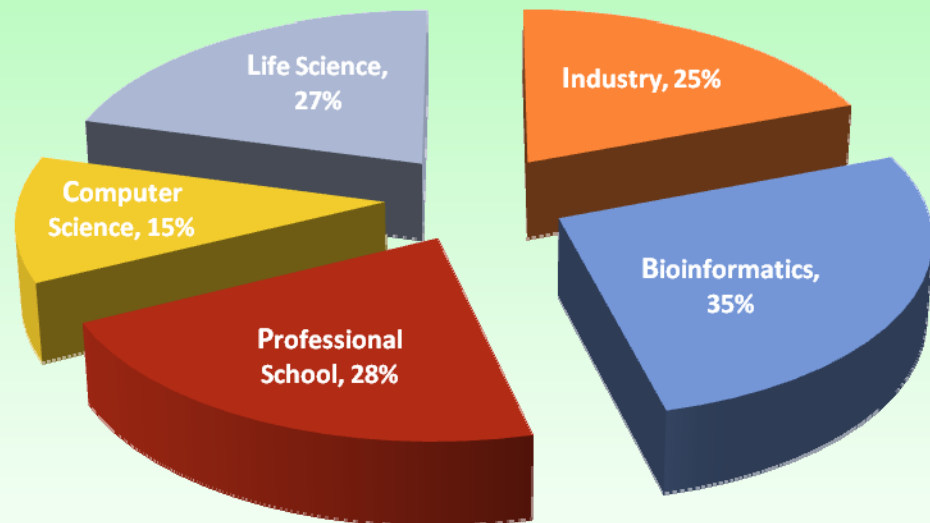
What can I do with this degree?

- **Graduate Schools (75%)**

- Bioinformatics
- Professional schools
- Life science
- Computer science

- **Industry (25%)**

- Bioinformatics
(pharmaceutical research)
- Computer science
- Life science



Bioinformatics Career-Outlook

Bioinformatics Career & Salary Outlook

Molecular modelers and bioinformatics software developers can obtain between \$60,000 and \$100,000.

The outlook for careers in bioinformatics is good, as the increasing trend in science is towards taking a multidisciplinary approach to solving biological problems. All of these career fields should see faster than average growth over the next five to seven years. However, competition for such jobs is expected to be high, and the best opportunities will be for candidates with post-graduate degrees.

Research Topics at Baylor

▪ Dr. Erich Baker

- *Shewanella* Knowledgebase: integrating technology across disciplines and between laboratories. (DoE)
- PTSD: managing data created by the center of excellence for traumatic brain injury. (DoD/VA)
- Ontological Discovery Environment: identifying empirically created ontologies from primary biological sources. (NIH)

▪ Dr. Young-Rae Cho

- Biological network modeling and analysis
- Genomic and proteomic functional characterization
- Disease-related pathway and drug target identification

Reference

- <http://www.ecs.baylor.edu/bioinformatics/>

School of Engineering Computer Science

Mastering Multiple Disciplines
A DECADE OF GROWTH IN BAYLOR BIOINFORMATICS

BIOINFORMATICS

BIOINFORMATICS FACULTY

EXPLORE BIOINFORMATICS

- About Us
- BSI Program
- Admissions
- Past and Present Students

Contact Information

Dr. Erich Baker
Phone: 254.710.7304
Erich_Baker@Baylor.edu

Michelle Aars
Phone: 254.710.3876
Michelle_Aars@baylor.edu

Mailing Address
One Bear Place #97356
Waco, TX 76798

Give Back to Baylor

Dr. Erich Baker started his education at the University of Notre Dame. He earned a B.A. in Political Science. He then continued his education receiving a M.S. in Biomedical Science from Barry University. Dr. Baker completed his education at the University of Tennessee at Knoxville where he earned his PhD in Comparative and Experimental Medicine.

Before coming to Baylor, he was an instructor at the University of Tennessee. In 2002, he joined the Baylor Department of Computer Science as an assistant professor in Bioinformatics.

Dr. Young-Rae Cho, an Assistant Professor in Computer Science, holds a B.S. in Material Science and Engineering from Yonsei University (1989), a B.A. in Computer and Information Science from Ohio State University (2002), a M.S. in Computer Science from the University of Illinois (2003) and a Ph.D. in Computer Science and Engineering from SUNY Buffalo in 2009.

ABOUT THE BIOINFORMATICS PROGRAM

WHAT IS BIOINFORMATICS?