

Bio-Computing

Young-Rae Cho, Ph.D.

Associate Professor

Division of Software / Division of Digital Healthcare

Yonsei University – Mirae Campus

Introduction to Bio-Computing



- ❑ Related Courses in US
 - Computational Biology (Undergraduate Junior or Senior Level)
 - Bioinformatics (Graduate Level)

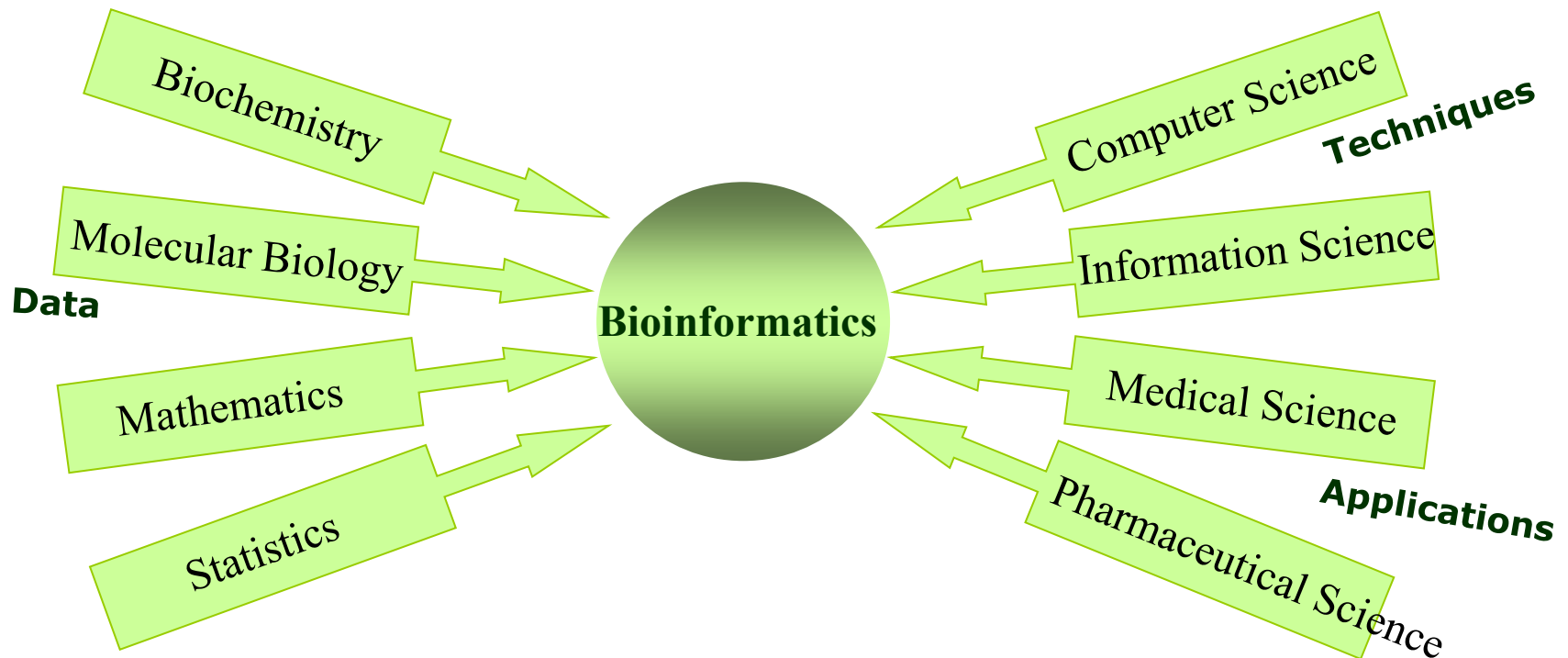
- ❑ Courses at Yonsei University – Mirae Campus
 - SWE4016/DHC3003 “Bio-Computing” (Undergraduate Junior or Senior Level)
 - ITD6043 “Bio-Data Mining” (Graduate Level)

Bioinformatics ?



- ❑ Interdisciplinary Research Area

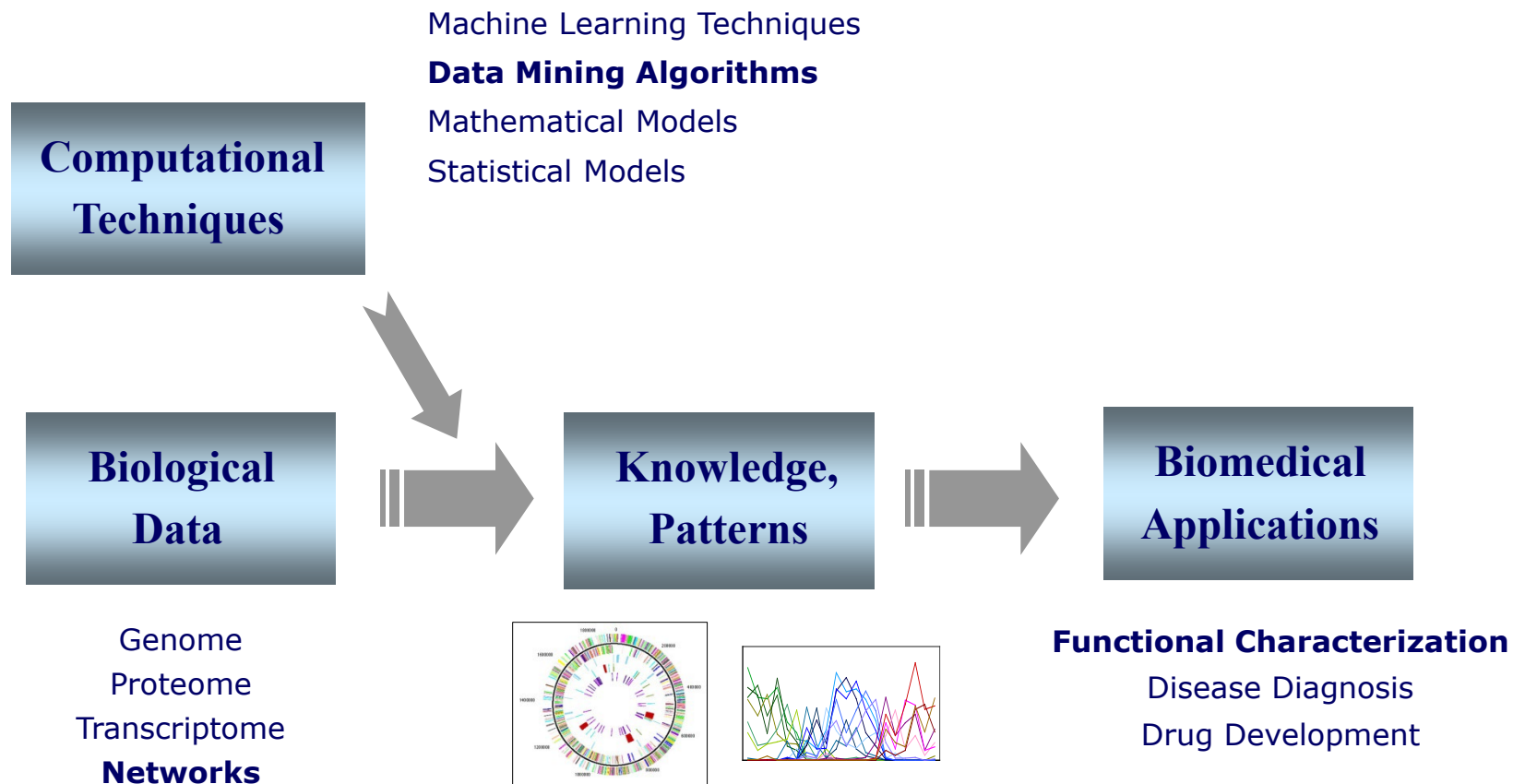
to *manage* and *analyze* biological data using computational techniques



Bioinformatics ?



□ Bioinformatics Research Process



Bioinformatics Milestone

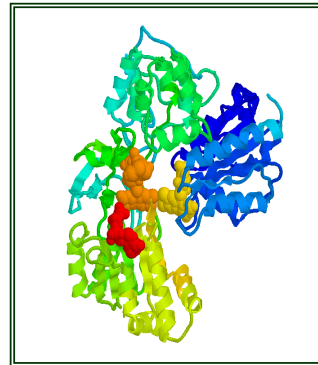
Computational Biology

Stage 1.
Sequence Analysis

Residue - 16024	(Sequenced Strand)	Sequence Length: 1122
16024	TTCITTCATG GGGAGCGA TTTGGGTACC ACCCAAGTAT TGACTACCC	
16074	ATCAACACCC GCTATGTATT TGGTACATTA CTGGACAGCA CCATGATAT	
16124	TGTTKGGTAC CATAAATAC TGGCCACCTG TATGTCATAG AAKCCGATC	
16174	CAKATCAAAA CCCCCTCCCC ATGCTTACAA GCAAGTACAG CAATCAACCC	
16224	TCBACTATCA CACATCAACT GCACTTCCAA AGCAACCCCT CACCCACTAG	
16274	GATACACACA AECTTACCA CCGTTACAG TACATATGAC ATAAAGCAT	
16324	TTTCCGTACA TGGACATTA KAGTCABAAT CCTTCTGTL CCATGGATG	
16374	ACCCCCCTCA GATAGGGGTC CCTTGACCAC CATCTTCGTT GAATCAATA	
16424	TCCCGCACAA GAGTGTACT CTCCTCGCTC CGGGCCATA ACCTTGGGG	
16474	GTAGTTAAG TGAUCTGTAT GCGACATCTG GTTCTTACTT CAGGTGATA	
16524	AAGCCTAAT AGCCACACG TTCCCTTAA ATAGACATC ACGTGGATC	
16574	ACHGGTCTAT CACCCATTA ACCACTCAGG GGAGCTTCC ATGCAITGG	
16624	TATTTTCTGC TGGGGGATGT GCACCGATA GCATCTCGAG BSKTGGAGC	
16674	CGGAGCCCC TATGTCGAG TATCTGCTT TGAATCTGC CTACCTCAT	
16724	TATTTATGC ACCTAGTTC AATATTACAG GCGAACATC TTAATAAGT	
16774	GTCTTAATTA ATTAATGCTT GTAGGACATA AATAATACAG TGTATGCTC	
16824	GCACAGCAC TTTCCAGCA GACATATATA CAAAAAATTI CGGCAACCC	
16874	CCCCCTCCCC CGGTTCTGGC CACAGCACTI AAMACATCT CTGCCAACC	
16924	CCAAAACAAA AGAACCCCTAA CACCAGCTTA ACCAGATTTC AAATTTTATC	
16974	TTTTGGGGGT ATGACATTTT ACHGATCAG CCCCAGTAA CACATTTT	
17024	TCCCTCCCA CTCCCATACT ACTAATCTCA TCAATACAC CCCCCECAT	
17074	CCTACCCAGC ACACACACAC CGCTGCTAAC CCCATACCCC GAACCAACCA	
17124	AACCCCAAG AGACCCCA CA	



Stage 2.
Structure Analysis



- DNA sequencing
- Homolog search
- Motif finding

- Protein folding
- Homolog search
- Binding site prediction

Human Genome Project

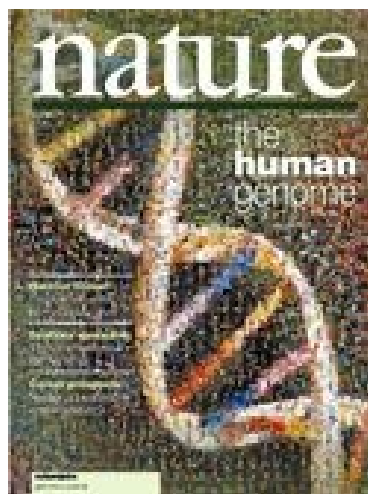


□ Goal

- Identification of complete human genome
- Mapping the gene from a functional standpoint

□ History

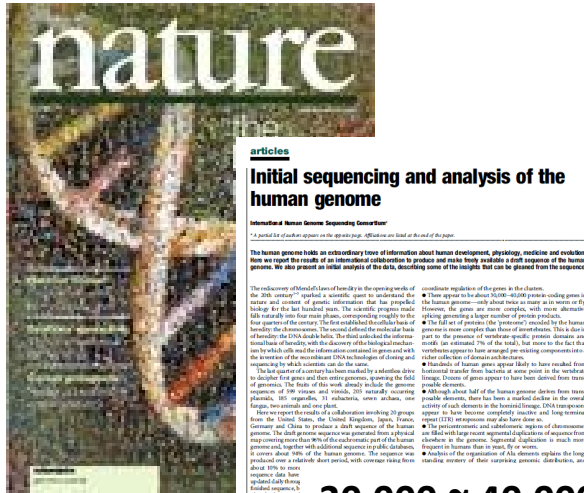
- Public project performed by Human Genome Project Consortium
- Private project performed by Celera Genomics, Craig Venter
- Both published final reports (called initial drafts) in 2001





The Number of Human Genes?

Human Genome Project Consortium



Initial sequencing and analysis of the human genome

International Human Genome Sequencing Consortium
A journal for authors appears on the open page. (Click on an article for the page.)

The human genome holds an extraordinary trove of information about human development, physiology, medicine and evolution. Here we report the results of an international collaboration to produce and make freely available a draft sequence of the human genome. We also present an initial analysis of the data, including some of the insights that can be gleaned from this sequence.

The recovery of 3 billion base pairs of sequence data in the 2001 project... The human genome holds an extraordinary trove of information about human development, physiology, medicine and evolution. Here we report the results of an international collaboration to produce and make freely available a draft sequence of the human genome. We also present an initial analysis of the data, including some of the insights that can be gleaned from this sequence.

30,000 ~ 40,000 genes

Celera Genomics



The Sequence of the Human Genome

J. Craig Venter,¹ Mark D. Adams,¹ Eugene W. Myers,¹ Peter W. Li,¹ Richard J. Mural,¹ George C. Sutton,¹ Hamilton O. Smith,¹ Mark Yandell,¹ Cheryl A. Evans,¹ Robert A. Holt,¹ Suzanne D. George,¹ Peter Annaraschi,¹ Richard M. Belloni,¹ Daniel M. Hays,¹ Jennifer Russo Wortman,¹ Qing Chang,¹ Chongjin O. Kohr,¹ Xiangqun H. Zhang,¹ Lin Chen,¹ Maria Skoultsi,¹ Gangadhara Subramanian,¹ Paul D. Thomas,¹ Jinghui Zhang,¹ George J. Gabor Miklos,¹ Catherine Nelson,¹ Samuel Brode,¹ Andrew G. Clark,¹ Joe Nadeau,¹ Victor A. McKusick,¹ Norton Zinder,¹ Anand J. Reddy,¹ Richard J. Roberts,¹ Fred Simeon,¹ Carolyn Slayman,¹ Michael Hunkapiller,¹ Randall Bolanos,¹ Arthur Delcher,¹ Ian Dew,¹ Daniel Fezza,¹ Michael Flanagan,¹ Liliana Flores,¹ Aaron Halpern,¹ Sidharth Hanamkhalil,¹ Said Kazir,¹ Samuel Levy,¹ Clark Mobary,¹ Ernst Moller,¹ Karlo Rimington,¹ Jane Abu-Threideh,¹ Ellen Beasley,¹ Kendra Bickel,¹ Vikas Bhatnagar,¹ Rhonda Braxton,¹ Michele Carroll,¹ Louise Chantanoonwong,¹ Ruzana Cherkas,¹ Fabir Chaturvedi,¹ Zhenming Dong,¹ Valentin D. Francesco,¹ Patrick Dunn,¹ Karen E. Elisei,¹ Carlos Evangelista,¹ Andrew S. Gershenson,¹ Weidong Guo,¹ Wangming Guo,¹ Fangsheng Gong,¹ Zhiping Guo,¹ Ping Guo,¹ Thomas J. Hellman,¹ Haoran E. Higgins,¹ Bai-Du Ji,¹ Zhanyi Ke,¹ Karen A. Ketchum,¹ Zhongyuan Li,¹ Tingting Li,¹ Zhang Li,¹ Jiyun Li,¹ Yong Liang,¹ Xiangping Lin,¹ Yu-Li Genseng Y. Markovits,¹ Natalia Mikhaleva,¹ Helen M. Moore,¹ Adalberto Ojeda,¹ X. Nish,¹ Valdear A. Nery,¹ Anne Hendrix,¹ Deborah Nankervis,¹ Douglas B. Rock,¹ Steven Sabing,¹ Wei Shen,¹ Baoping Shen,¹ Jian-Yun Wang,¹ Xian-Yun Wang,¹ Alan Wang,¹ Xin Wang,¹ Jian Wang,¹ Ming-Hui Wu,¹ Bao Wides,¹ Chuanlin Xia,¹ Chunhua Yan,¹ Alton Yan,¹ Jena Ye,¹ Ming Zhang,¹ Weiqing Zhang,¹ Hongyong Zhang,¹ Qi Zhang,¹ Liandong Zhang,¹ Fei Zhang,¹ Wenyang Zhang,¹ Shiyong C. Zhu,¹ Shuying Zhu,¹ Dennis Gilbert,¹ Susanna Baumhueter,¹ Gene Spier,¹ Christine Carter,¹ Anibal Corvalan,¹ Trevor Woodage,¹ Ferris Ali,¹ Hajira Ali,¹ Adarsh Anand,¹ Dariusz Badwin,¹ Holly Bales,¹ Mary Bamstead,¹ Ian Barrow,¹ Karen Benson,¹ Dana Bissim,¹ Amy Carver,¹ Angela Center,¹ Hong-Gui Cheng,¹ Liu Curry,¹ Steve Danaher,¹ Lionel Davenport,¹ Raymond Deshler,¹ Suzanne Dietz,¹ Kristina Dowland,¹ Lisa Dooy,¹ Steven Ferreira,¹ Neha Garg,¹ Andre Glandermans,¹ Bill Hart,¹ Jason Haynes,¹ Charles Haynes,¹ Cheryl Hebert,¹ Susanna Hedges,¹ Dusan Hostin,¹ Jarrett Housh,¹ Timothy Howland,¹ Chloeyre Bregman,¹ Jeffrey Johnson,¹ Francis Kahana,¹ Lindsay Kline,¹ Shaoh Kolonel,¹ Amy Loo,¹ Felicia Miller,¹ David Mui,¹ Steven McCawley,¹ Tina Mihalich,¹ by McMillan,¹ Mei Mui,¹ Linda Moy,¹ Brian Murphy,¹ Kathleen Poulos,¹ Michael Rhee,¹ Peter Rhee,¹ Vincent Rhee,¹ Loretta Poulos,¹ Kathleen Poulos,¹

26,588 + 12,000 = 38,588 genes

Conclusion

- Nobody can confirm how many genes human has!

Bioinformatics Milestone



Computational Biology

Functional Genomics

Systems Biology

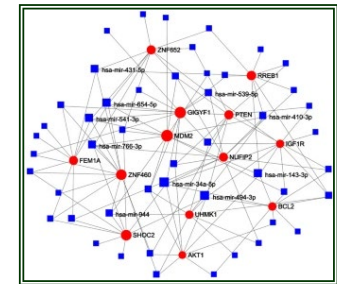
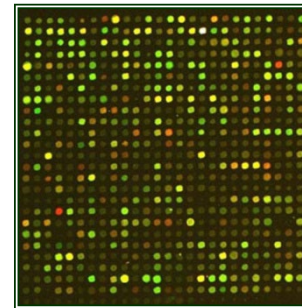
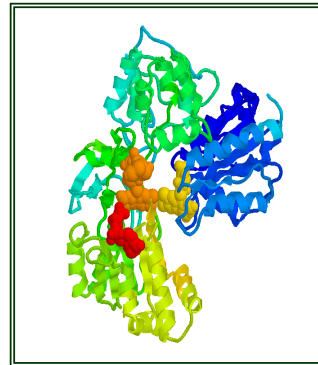
Stage 1.
Sequence Analysis

Stage 2.
Structure Analysis

Stage 3.
Genome Analysis

Stage 4.
Network Analysis

Residue: 16024	(Sequenced Strand)	Sequence Length: 1122
16024	TTCTTTCATG GGGAGCAGA TTGGGTACC ACCCAAGTAT TGACTACCC	
16074	ATCAACACC GCTATGATTT TGGTACATTA CTGCACACA CCATGATAT	
16124	TTTACGGTAC CATAAATATC TGCCACCTG TATGACATAA AAACCCATC	
16174	CACATCAAAA CCCCCTCCC ATGCTTACA GCAAGTACAG CAATCAACCC	
16224	TCACATACA CACATCAACT GCACATCAA GGCACCCCT CACCCACTAG	
16274	GATACACACA AACTCAACA CCCCACACG TACATATATA ATAAAGCAT	
16324	TTCCGTACA TGCACATTA CAGTCAAATC CCTCTGCTC CCGATGGATG	
16374	ACCCCCCA GATAGGGTTC CCTTGACCAC CATECTCGT GAATCAATA	
16424	TCCCGCACAA GAGTGGTACT CTCCTCGCTC CGGGCCATA ACACCTGGGG	
16474	GTAGTAAAG TGACTGTAT CCGACATCTG GTTCTTACT CAGGTGATA	
16524	AAGCTAAAT AGCCACACG TTCCCTTAA ATAGACATC ACGTGGATC	
16574	ACAGGCTAT CACCTATTA ACCACTCAGG GGAGCTTCC ATGCATTGG	
16624	TATTTTCTG TGGGGGATCT GCACCGATA GCATTCTGG GCGTGGAGC	
16674	CGGACACCC TATGTCGAG TATTTGCTT TGATCTGCT CTACCTCAT	
16724	TATTTATGC ACCTACGTT CATTATTACG GCGACATAC TTAATAAGT	
16774	GCTTAAATTA ATTAATGCTT GTAGGACATA AATAATACAA TGTATGCTC	
16824	GACACACAC TTTCACACA GAGTATATA CAAAAATTC CCGACAAAC	
16874	CCCCCTCCC CGCTTCTGG CACAGACTT AAACACATC CTGCCAACC	
16924	CCAAAACAA AGAACCTTAA CACACGCTA ACCAGATTTC AAATTTTATC	
16974	TTTTGGGCT ATGACATTTT ACGGTGAC CCCCACATA CAGTATATT	
17024	TCCCTCCCA CTCCCATCT ACTAATCTCA TCAATACAC CCCCCECAT	
17074	CCTACCCAG ACACACACAC CGCTGCTAAC CCCATACCC GAACCAACCA	
17124	AACCCAAAG ACACCCCCA CA	



- DNA sequencing
- Homolog search
- Motif finding

- Protein folding
- Homolog search
- Binding site prediction

- Function prediction
- Gene clustering
- Gene classification

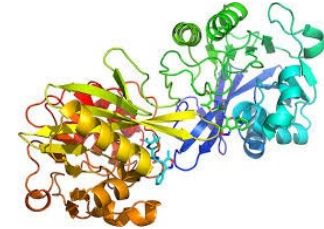
- Network modeling
- Module finding
- Pathway prediction

Computational Network Biology - Data



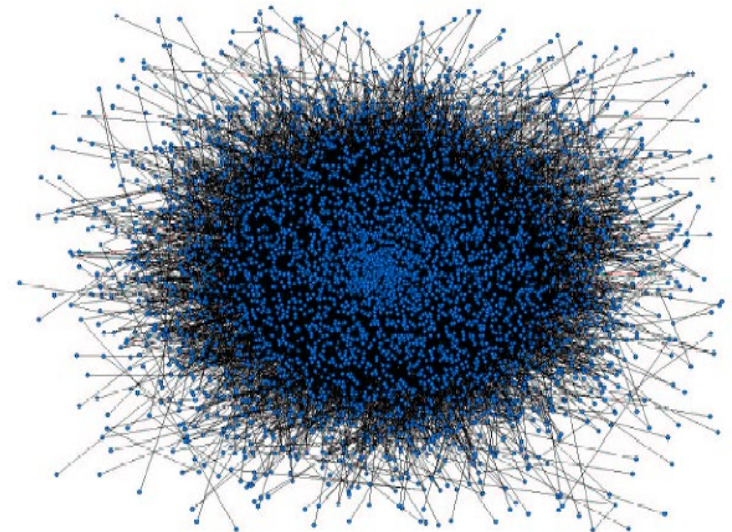
❑ Interactome

- Genome-wide PPIs (Protein-Protein Interactions)
- Problem?
 - Large scale & Unreliability
- Demand?
 - Computational, integrative approaches



❑ PPI Networks

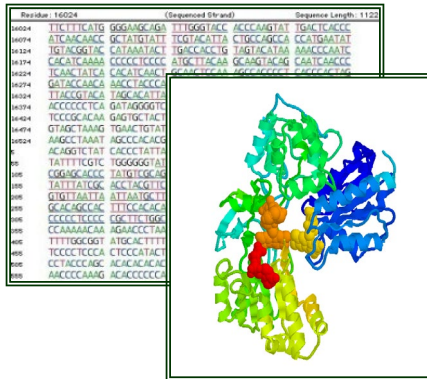
- Undirected unweighed graph, $G(V,E)$
- Problem?
 - Complex connectivity
- Demand?
 - Computational, systematic approaches



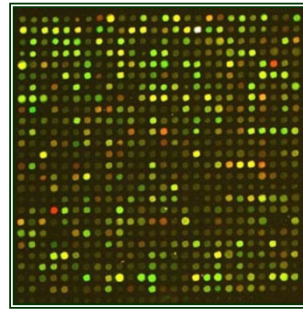
From Bioinformatics To Biomedical Informatics



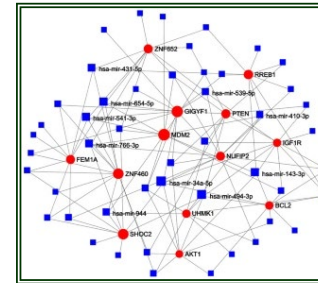
Stage 1/2. Sequence & Structure Analysis



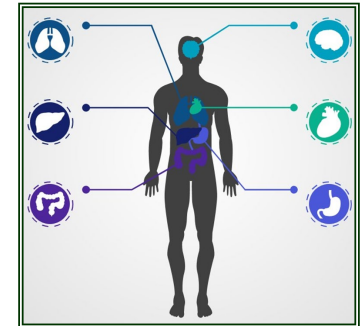
Stage 3. Genome Analysis



Stage 4. Network Analysis



Stage 5. Medical Data Analysis



- Function prediction
- Gene clustering
- Gene classification

- Network modeling
- Module finding
- Pathway prediction

- Disease Diagnosis
- Drug Development

Personalized Medicine / Precision Medicine



❑ Generalized Concept

- Customized healthcare including medical treatment and prevention
- Background: All individuals have genetic variation

❑ Use of Big Data

- Genomic data and patients' medical record data

❑ Support by Advanced Technology

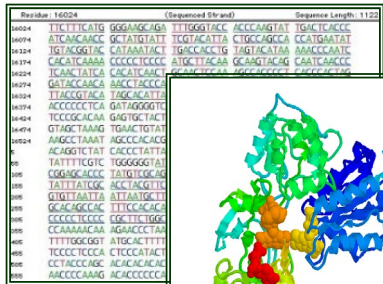
- Machine learning, deep learning, and artificial intelligence (AI) techniques



Scope of Bio-Computing

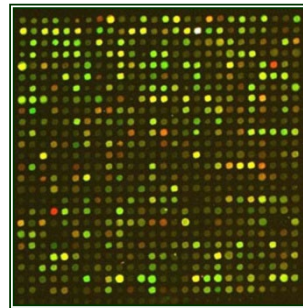


Stage 1/2. Sequence & Structure Analysis



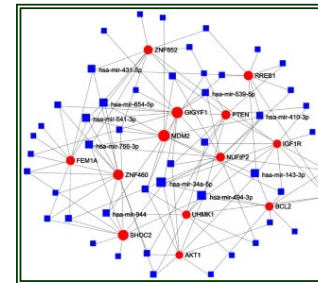
- DNA sequencing
- Homolog search
- Motif finding
- Protein folding
- Binding site prediction

Stage 3. Genome Analysis



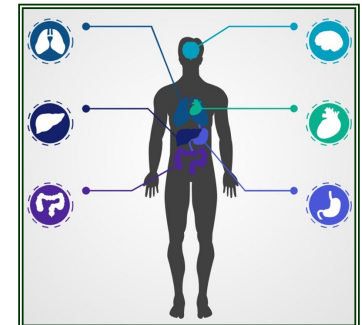
- Function prediction
- Gene clustering
- Gene classification

Stage 4. Network Analysis



- Network modeling
- Module finding
- Pathway prediction

Stage 5. Medical Data Analysis



- Disease Diagnosis
- Drug Development

Subjects of Bio-Computing



❑ Sequence Analysis

- Pairwise Sequence Alignment
- Multiple Sequence Alignment
- Pattern Matching and Finding
- Phylogenetic Analysis

❑ Genome Analysis

- Gene Clustering
- Gene Classification

❑ Network Analysis

- Network Modelling
- Module Finding

