Computational Science, Bioinformatics and Biomedical Informatics

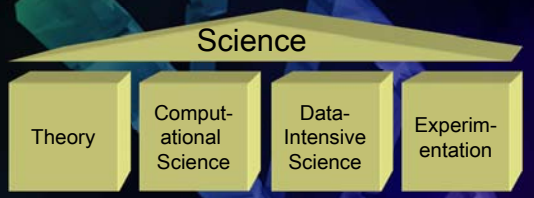
Tu Bao Ho
School of Knowledge Science
Japan Advanced Institute of Science and Technology

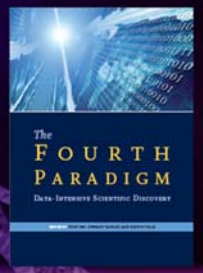
Content

- Computational science
- Bioinformatics and Biomedical informatics
- Examples in hepatitis study

A number of slides are adapted from
- American Medical Informatics Association: www.amia.org
- Talk on Biomedical informatics, Kun Huang
- Talk of Dr. T. V. Rao MD
- Talk of Eleftherios P. Diamandis

Science: Two, three or four legs?





CACM, Dec. 2010

CACM, Sep. 2010

PITAC report: ''Computational Science: Ensuring America's Competitiveness''

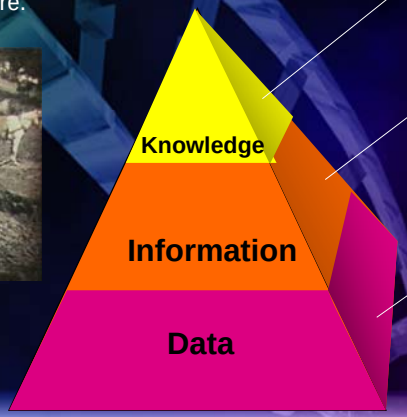
1. A Wake-up Call: The Challenges to U.S. Preeminence and Competitiveness
2. Medieval or Modern? Research and Education Structures for the 21st Century
3. Multi-decade Roadmap for Computational Science
4. Sustained Infrastructure for Discovery and Competitiveness
5. Research and Development Challenges
6. Appendices

PITAC: President's Information Technology Advisory Committee.

(24 leaders in industry and academia in the 6 of them in Computational Science Subcommittee.)

Data, Information and Knowledge

Metaphor:
Data: rock;
knowledge: ore.
Miner?



Is the size and staffs of Chiba University hospital appropriate for such amount of patients?

Average of number of patients each hour, each day, each week, each at Chiba Univ. hospital.

Number of patients counted at Chiba Univ. hospital by hours, by days of the week, by months.

Computer science vs. computational science

- **Computer science**: Also commonly understood as informatics or information technology, e.g., processing of information by computer.
- Science about producing hardware (computer) and software (programs) for different usage purpose.

- **Computational science**: Science about using computing and mathematics to do research in other sciences, e.g.
  - Computational physics
  - Computational finance
  - Computational linguistics
  - Computational biology

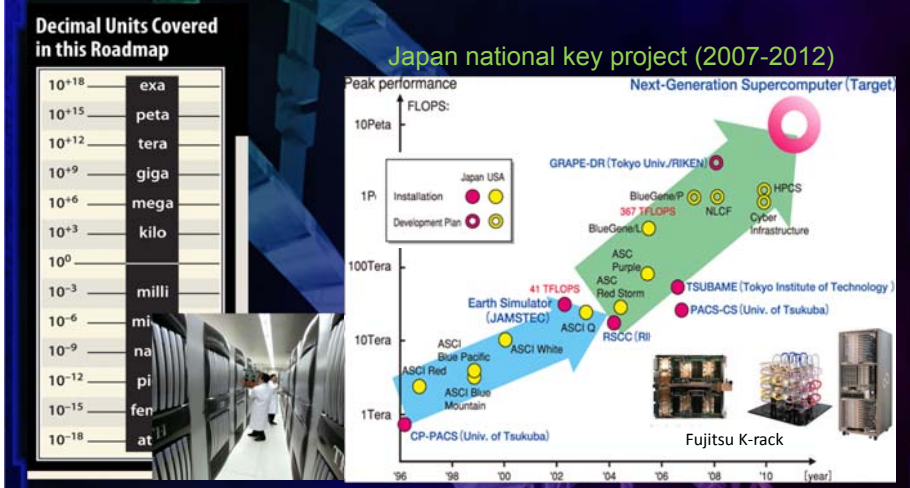
Computational Science

''is a rapidly growing multidisciplinary field that uses advanced computing capabilities to understand and solve complex problem''. Its fuses three distinct elements

- Algorithms and **modeling and simulation** software developed to solve science, engineering, and humanities problems
- **Computer and information science** that develops and optimizes the system hardware, software, networking and data management components needed to solve computationally demanding problems
- **The computing infrastructure** that supports both the science and engineering problem solving and the development computer and information science.

Book: Computational Science: Ensuring America's Competitiveness
PITAC: President's Information Technology Advisory Committee

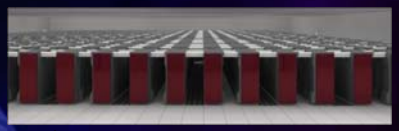
High performance computing competition



Tianhe-1A: 7,168 NVIDIA® Tesla™ M2050 GPUs and 14,336 CPUs (2.507 Peta flops)

High performance computing competition

- Japan's ''K computer'' (K is 10^{16} and large gateway), 800 computer racks with Fujitsu ultrafast CPUs, targeting by 2012 to 10 petaflop, (RIKEN's Advanced Institute for Computational Science)
- IBM's computers BlueGene and BlueWaters, targeting to 20 petaflop by 2012 (Lawrence Livermore National Laboratory).

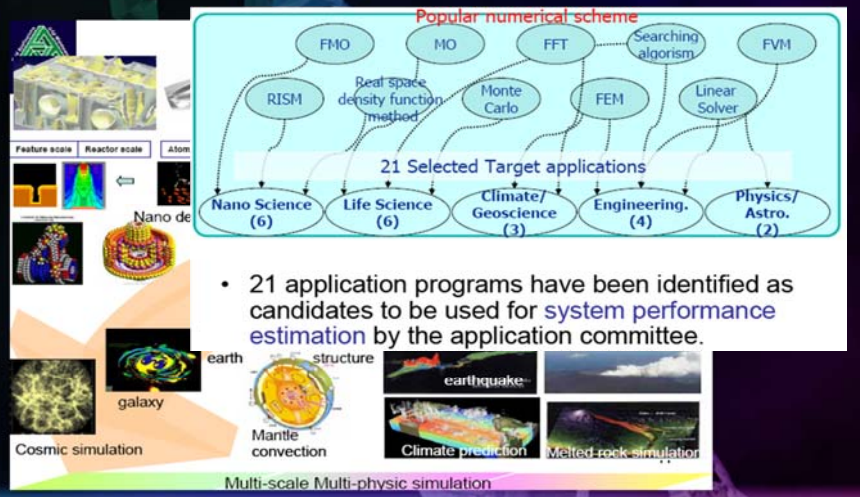


Japan's K computer

IBM BlueGene

http://www.fujitsu.com/global/news/pr/archives/month/2010/20100928-01.html (28.9.2010)
http://www.hightechnewstoday.com/nov-2010-high-tech-news/38-nov-23-2010-high-tech-news.shtml (Nov. 2010)

Computational science and supercomputers



Kennichi Miura, DEISA Symposium, 5.2007

Doctors are very accepting of technology!

Computer usage in medicine

- Earliest broad recognition of statistical issues in diagnosis and the potential role of computers occurred in the late 1950s
  - "Reasoning foundations in medical diagnosis": Classic article by Ledley and Lusted appeared in *Science* in 1959.
- Computers began to be applied in biomedicine in the 1960s
  - Most applications dealt with clinical issues, including diagnostic systems.

Computer usage in medicine

- "Computers in medicine" in the 1960s
  - First Federal grant review group
  - Most applications dealt with clinical issues
- No consistency in naming the field for many years
  - "Computer applications in medicine"
  - "Medical information sciences"
  - "Medical computer science"
- Emergence in the 1980s of a single, consistent name, derived from the European (French) term for computer science: informatique (informatics)
  - Medical Informatics

"Fundamental theorem" for using computers in medicine

Charles P. Friedman. J Am Med Inform Assoc. 2009;16:169 –170.

Related disciplines

- Informatics (computer science)
- Computational science
- Medical informatics (Clinical informatics)
- Public health informatics
- Biomedical informatics
- Molecular medicine
- Bioinformatics (computational biology)

Medical Informatics

Medical informatics is the intersection of computer science, computational science and health care. It deals with the resources, devices, and methods required to optimize the acquisition, storage, retrieval, and use of information in health and biomedicine.

Medical informatics is rapidly developing

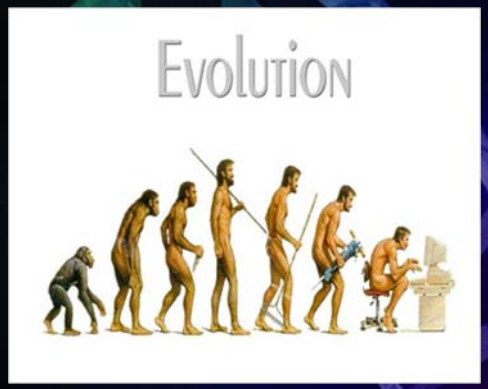
Medical informatics is the rapidly developing scientific field that deals with resources, devices and formalized methods for optimizing the storage, retrieval and management of biomedical information for problem solving and decision making.

Edward Shortliffe, M.D., Ph.D. What is medical informatics? Stanford University, 1995.

Examples of medical informatics areas

- Hospital information systems
  - Electronic medical records & medical vocabularies
  - Laboratory information systems
  - Pharmaceutical information systems
  - Radiological (imaging) information systems
  - Patient monitoring systems
- Clinical decision-support systems
  - Diagnosis/interpretation
  - Therapy/management

Are too slow adopting the change

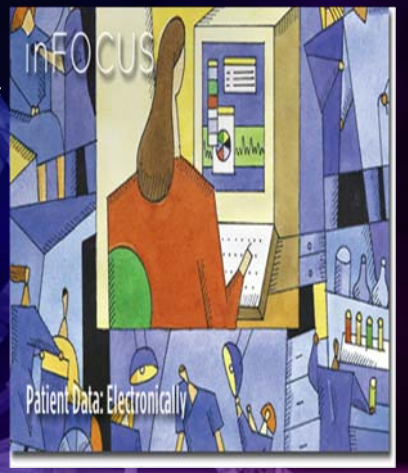


Medical schools have long recognized the need to revise their teaching methodology, but have been slow to change.

Patient data is the most important resource

As medical knowledge continues to expand rapidly with demands for more efficient coordination of patient data become paramount, and the pressures for improved practice and application of evidence based medicine increases, medical informatics will have increasing influence in our working lives as clinicians.

Clinical data vs. omics data!



What is biomedical informatics?

Biomedical informatics (BMI) is the interdisciplinary field that studies and pursues the effective uses of biomedical data, information, and knowledge for scientific inquiry, problem solving, and decision making, motivated by efforts to improve human health.

Source: www.amia.org

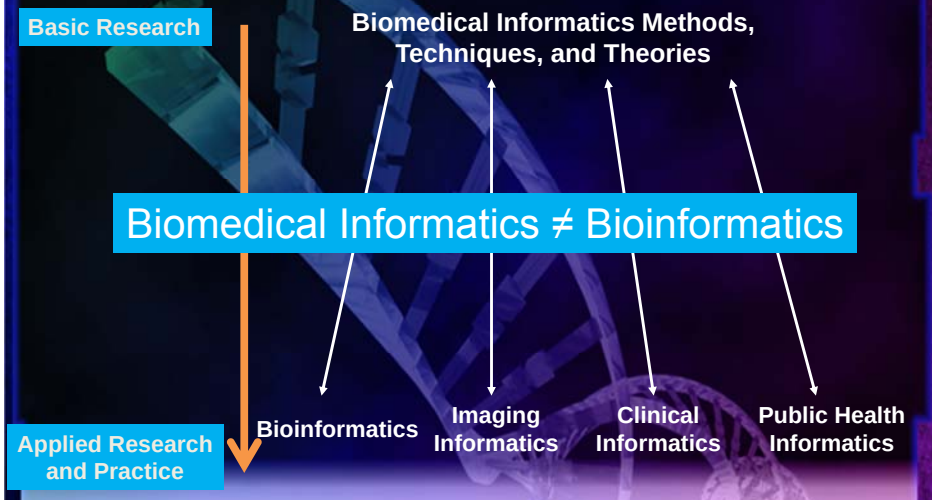
Biomedical informatics: Corollaries to the definition

1. BMI develops, studies and applies theories, methods and processes for the generation, storage, retrieval, use, and sharing of biomedical data, information, and knowledge.
2. BMI builds on computing, communication and information sciences and technologies and their application in biomedicine.

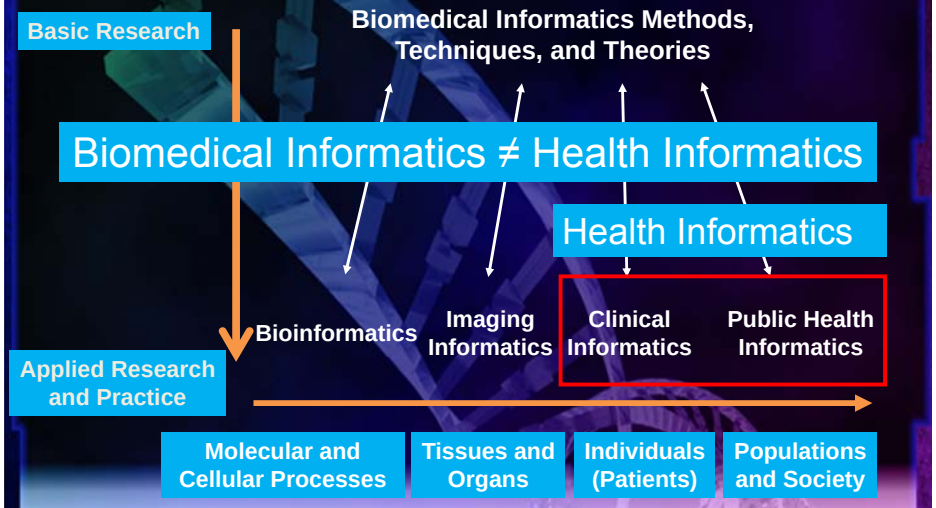
Biomedical informatics: Corollaries to the definition

3. BMI investigates and supports reasoning, modeling, simulation, experimentation and translation across the spectrum from molecules to populations, dealing with a variety of biological systems, bridging basic and clinical research and practice, and the healthcare enterprise.
4. BMI, recognizing that people are the ultimate users of biomedical information, draws upon the social and behavioral sciences to inform the design and evaluation of technical solutions and the evolution of complex economic, ethical, social, educational, and organizational systems

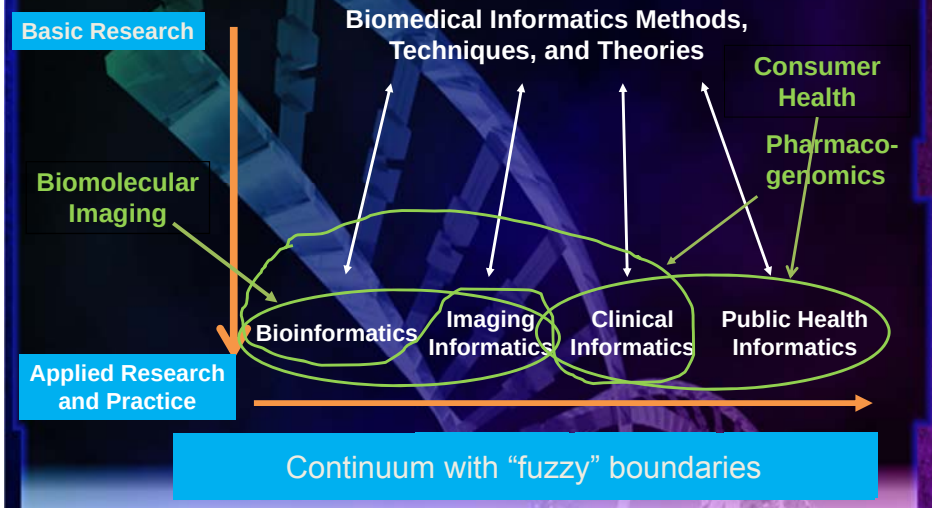
Biomedical informatics in perspective



Biomedical informatics in perspective



Biomedical informatics in perspective



KDD: Knowledge Discovery and Data Mining
知識発見とデータマイニング

"Data-driven discovery of models and patterns from massive observational data sets"

- Statistics, Inference
- Languages, Representations
- Data Management

→ Applications

Example: mining associations in market data
マーケット・バスケット分析(IBM)

Super market data

"Young men buy diaper and beer together"
「紙おむつを買う男性は缶ビールを一緒に買うことが多い」

(解釈:顧客像) 紙おむつを買うように頼まれた男性がついでに自分用の缶ビールを購入していた → 今後の陳列に活かすことのできる知識.

Text mining: a real example (Swanson,1997)

Extract pieces of evidence from article titles in the biomedical literature　生物医学文献タイトルからの科学的根拠の抽出

- ✓ “stress is associated with migraines” “ストレスは片頭痛を伴う”
- ✓ “stress can lead to loss of magnesium”
  “ストレスはマグネシウム損失の原因となる”
- ✓ “calcium channel blockers prevent some migraines”
  “カルシウム拮抗薬は片頭痛を予防することがある”
- ✓ “magnesium is a natural calcium channel blocker”
  “マグネシウムは天然のカルシウム拮抗薬である”

Induce a new hypothesis not in the literature by combining culled text fragments with human medical expertise
抜粋した文の断片を人間の医学専門知識を使って組合せ，文献にない新しい仮説を導き出す

- ✓ Magnesium deficiency may play a role in some kinds of migraine headache　マグネシウムはある種の片頭痛に関与するらしい

Data schemas vs. mining methods
データ･スキーマ vs. 学習手法

Types of data

- Flat data tables　表形式データ
- Relational databases　関係DB
- Temporal & spatial data
  時空間データ
- Transactional databases
  取引データ
- Multimedia data
  マルチメディアデータ
- Genome databases ゲノムデータ
- Materials science data
  材料データ
- Textual data　テキストデータ
- Web data　ウェブデータ
- etc.

Mining tasks and methods
マイニングの課題と手法

- **Classification/Prediction　分類/予測**
  - **Decision trees　決定木**
  - Neural networks　神経回路網
  - Rule induction　ルール帰納法
  - Support vector machines　SVM
  - Hidden Markov Model 隠れマルコフ
  - etc.
- **Description　記述**
  - Association analysis　相関分析
  - Clustering　クラスタリング
  - Summarization　要約
  - etc.

The KDD process
知識発見とデータマイニングのプロセス

a step in the KDD process consisting of methods that produce useful patterns or models from the data
データから有用なパターンやモデルを生成する手法からなるステップ

Maybe 70-90% of effort and cost in KDD 全体の70-90%の労力を要するステップ

5. Putting the results in practical use
   結果を実践に用いる
4. Interpret and evaluate discovered knowledge
   発見した知識を解釈し評価する
3. Data mining
   Extract Patterns/Models
   パターン･モデルを抽出する
2. Collect and preprocess data
   データを収集し前処理する
1. Understand the domain and define problems
   領域を理解し問題を定義する

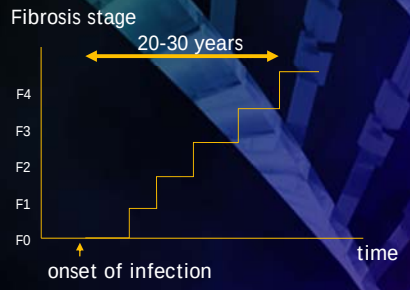
KDD is inherently interactive and iterative
知識発見の本質は繰り返しとインタラクション

(In many cases, viewed KDD as data mining)

Case study: Hepatitis

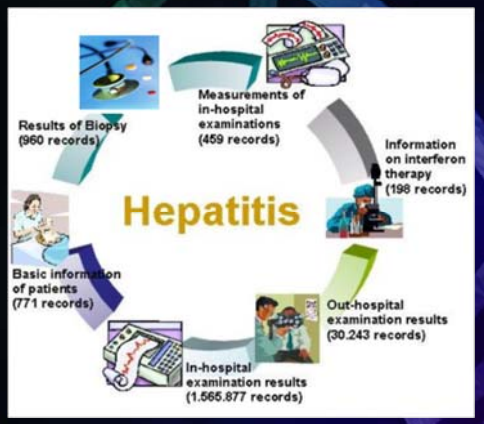
LC: liver cirrhosis
HCC: hepatocellular carcinoma

Hepatitis → LC → HCC



The natural course of hepatitis

The hepatitis dataset



- Temporal relational database (Chiba Univ. Hospital)
- Patient's data contain about 983 tests taken in different periods
  - ➔ varying from several weeks to twenty years
  - ➔ irregular time-stamped points

Example of the hepatitis dataset

Table 1. Part of integrated table of temporal data

MID	Date	Sex	IFN	GOT	GPT	ALB	…
1	19810219	M	n	55	65	5.4	…
1	19810316	M	n	54	87	5.2	…
1	19820313	M	n	47	64	5.2	…
…	…	…	…	…	…	…	…
1	20010108	M	y	68	100	5.5	…
1	20010510	M	y	57	93	5.1	…
2	19911021	F	n	54	82	4.5	…
2	19911118	F	n	77	114	4.4	…
…	…	…	…	…	…	…	…

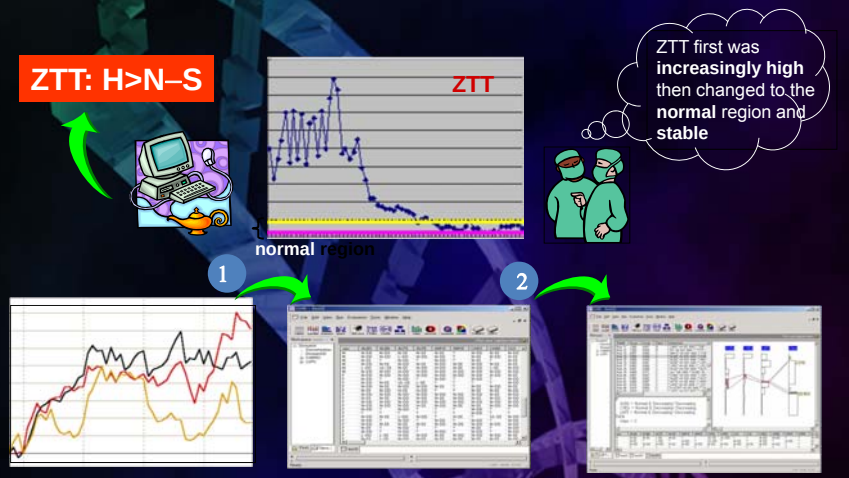
Research problems

P1. Differences in temporal patterns between hepatitis B and C? (HBV, HCV)

P2. Evaluate whether laboratory examinations can be used to estimate the stage of liver fibrosis? (F0, F1, F2, F3, F4)

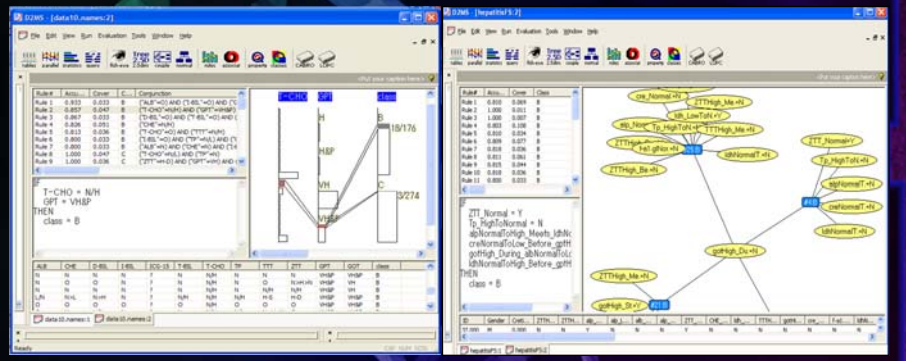
P3. Evaluate whether the interferon therapy is effective or not? (Response, Partial response, Aggravation, No response)

Our solution: temporal abstraction



Two methods: APE (abstraction pattern extraction) and TRE (temporal relation extraction)

System D2MS



C:\Users\HO TU BAO\Desktop\D2MS.lnk

C:\Users\HO TU BAO\Desktop\DBMS-noN.lnk

Lessons learned

1. Understanding the domain and determining the target are the primary factor to failure or success.
2. Pre-processing takes more than 90% of time and effort. "It took 40 years to collect data, months to preprocess data, minutes to learn from data, and hours/days to evaluates obtained results".
3. The collaboration between data miners and domain experts is the most decisive factor.
4. Different views on interestingness of discovered patterns are the main reason of un-satisfaction.
5. Model selection requires the active participation of users, and domain knowledge in mining is crucial.

What is bioinformatics?

- The collection, classification, storage, and analysis of biochemical and biological information using computers especially as applied in molecular genetics and genomics*.
- The application of math and computing to solve problems in biology.

Bioinformatics ~ Computational biology

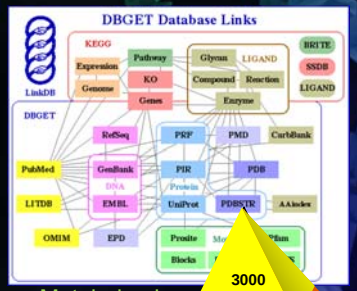
* *Merriam-Webster's Medical Dictionary, © 2002 Merriam-Webster, Inc.*

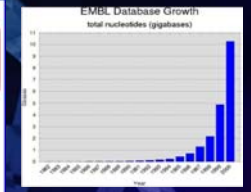
Bioinformatics is about biological data

- Nucleotide – DNA, RNA, …
- Genome – Sequences, chromosomes, expressed data, …
- Protein – Sequences, 3-D structure, interaction, …
- System – Gene network, protein network, TFs, …
- Other – Microarray, images, lab records, journals, literatures, …

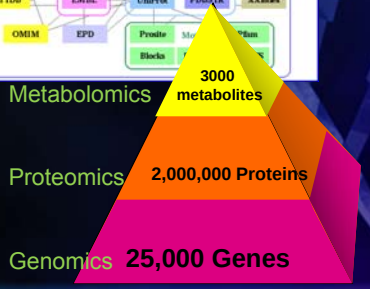
The goal is to understand how the biological system works.

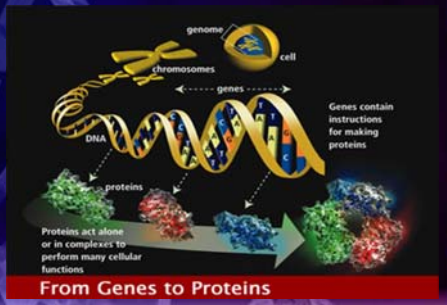
Explosion of biological data





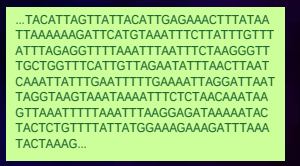
10,267,507,282 bases in 9,092,760 records.

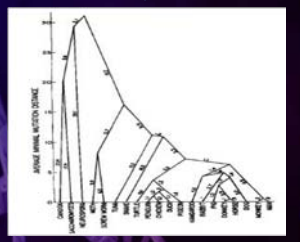




From data to knowledge

- **Data**
  - Nucleotide – DNA, RNA, …
  - Genome – Sequences, chromosomes, expressed data, …
  - Protein – Sequences, 3-D structure, interaction, …
  - System – Gene network, protein network, TFs, …
  - Other – Masspec, microarray, images, lab records, journals, literatures, …
- **Knowledge**
  - Genotype
  - Phenotype
  - Genotype-Phenotype relationship
  - SNPs
  - Pathways
  - Drug targets





Getting data is "easy", extracting knowledge is hard!

Computer is intelligent

- **Pros**
  - Repeated work
  - Accurate storage
  - Precise computation
  - Fast communication
  - …
- **Cons**
  - Cannot generalize
  - No real intelligence
  - …

The results must be reviewed and validated by biologists. In addition, biologists must have some understanding of how computer processes data (algorithms).

The … omics

Level of analysis	Definition	Status	Method of analysis
Genome	Complete set of genes of an organism or its organelles	Context-dependent (modifications to the yeast genome may be made with exquisite precision)	Systematic DNA sequencing
Transcriptome	Complete set of messenger RNA molecules present in a cell, tissue or organ	Context-dependent (the complement of messenger RNAs varies with changes in physiology, development or pathology)	Hybridizations arrays SAGE High-throughput Northern analysis
Proteome	Complete set of protein molecules present in a cell, tissue or organ	Context-dependent	Two-dimensional gel electrophoresis, peptide mass fingerprinting
Metabolome	Complete set of metabolites (low-molecular-weights intermediates) in a cell, tissue or organ	Context-dependent	Infre-red Spectroscopy Mass spectroscopy NMR spectrometry

Elements of bioinformatics

- **Genomics** is a discipline in genetics concerning the study of the genomes of organisms.
- **Transcriptomics (functional genetics)** is the process of creating a complementary RNA copy of a sequence of DNA.
- **Proteomics** is the large-scale study of proteins, particularly their structures and functions.
- **Metabolomics** is the scientific study of chemical processes involving metabolites.

How to extract knowledge?

Computational tools

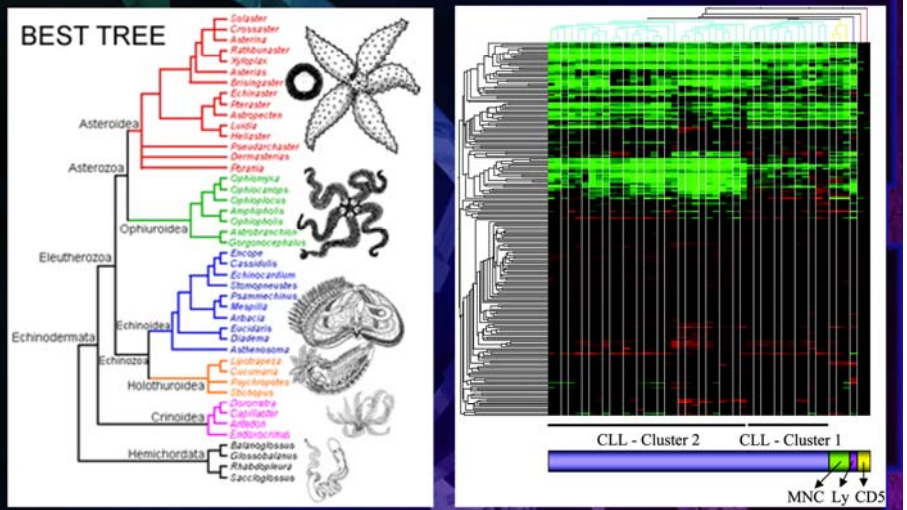
- Building the databases
- Perform analysis/extract features
- Data fusion/Integration
- Data mining/Statistical learning
- Visualization/representation

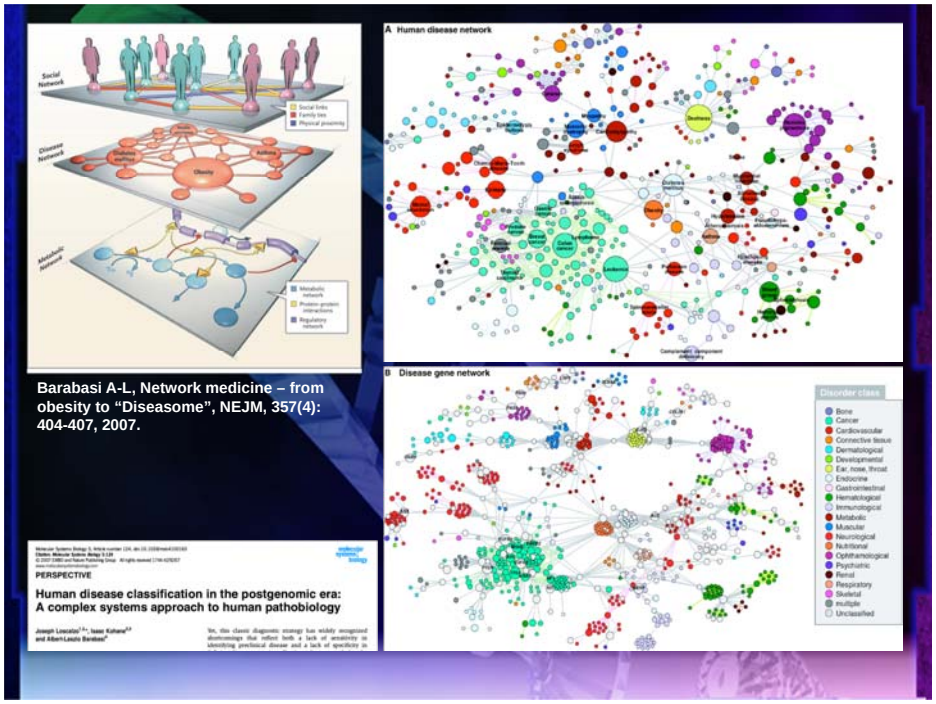
Biological information and knowledge

Statistics vs. data mining

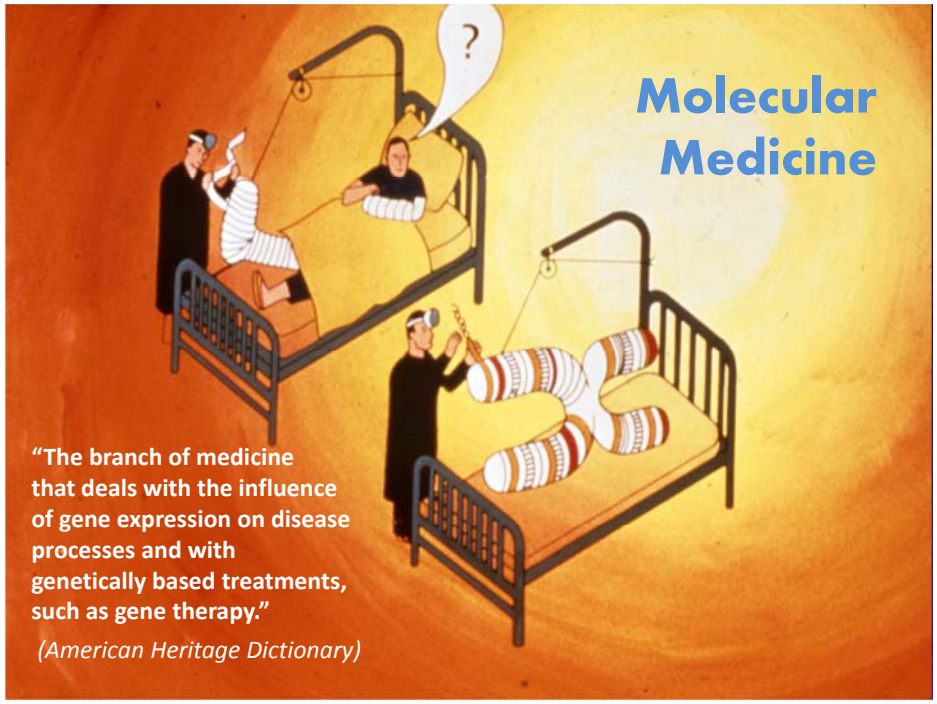
- **Statistics** provides principles and methodology for designing the process of
  - Data Collection
  - Summarizing and Interpreting the data
  - Drawing conclusions or generalities
- **Data mining**
  - Finding knowledge from data
  - Strongly based on statistics, especially modern multivariate statistics
  - Also based all other disciplines
  - Motivated by real-world problems

How to extract knowledge?





Barabasi A-L, Network medicine – from obesity to "Diseasome", NEJM, 357(4): 404-407, 2007.



Molecular medicine

- **Molecular medicine** is a broad field, where physical, chemical, biological and medical techniques are used to describe molecular structures and mechanisms, identify fundamental molecular and genetic errors of disease, and to develop molecular interventions to correct them.
- The molecular medicine perspective emphasizes cellular and molecular phenomena and interventions rather than the previous conceptual and observational focus on patients and their organs.

Trends in Molecular Medicine

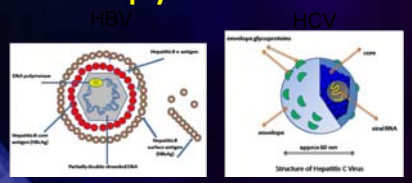
Our work in biomedicine

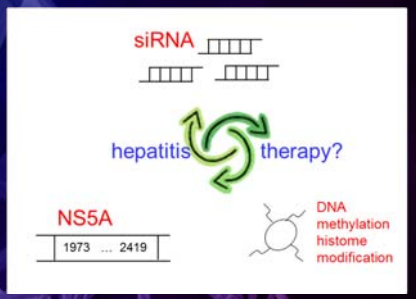
- Computational medicine
  - Mining stomach cancer data, Tokyo Cancer Center (1999-2003)
  - Mining hepatitis data, Chiba University hospital (2001-2005)
  - Hepatitis study (2007-)
- Computational biology
  - Transcriptional regulation
  - Epigenetics
  - Protein-protein interactions
  - miRNA
  - Metabolomics

4 PhD graduated, 4 in progress

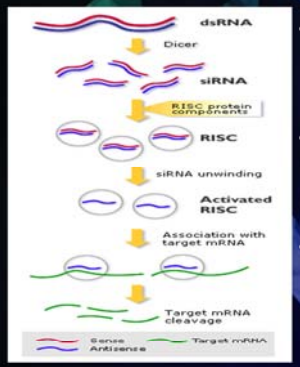
HCV NS5A and IFN/RBV therapy

- 300 and 170 millions of carriers of HBV and HCV worldwide, respectively.
- Lead to liver cirrhosis & HCC
- Molecular mechanisms of hepatitis pathogenesis and hepatitis therapies
- 50% SVR for IFN/RBV
- Why and why not hepatitis viruses disappear after the treatment?





RNA interference (RNAi) and hepatitis



- A mechanism wherein small RNAs, esp. miRNA and siRNA, control the expression of genes.
- RNAi target to HBV and HCV genes to inhibit their replication or host genes required for their replication.
- How to select appropriate siRNA molecules that have satisfactory silencing capabilities or minimum off-target effect and maximum knockdown efficiency?

Fire, A., Mello, C., Nature 391, 1998 (Nobel Prize 2006)

Problem of siRNA selection

ACGCCGU 0.65
UAUUACG 0.85
ACACGGU 0.33
GGCUACC 0.90
AUUAUUC 0.34
CUAUGGA 0.51
AAGCGUA 0.47
UACCGGU 0.55

Set of siRNA with known score of knockdown efficacy (about 5000)

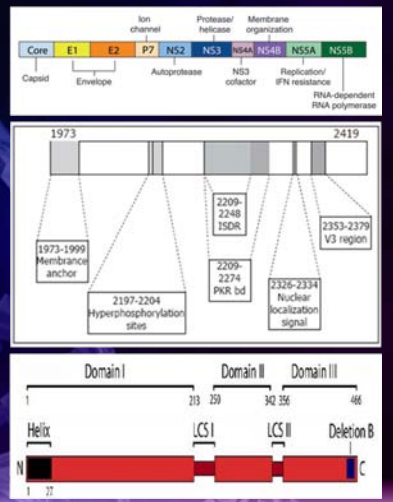
Position/Nucleotide	A	C	G	U
17	C> A> G	A >U> C		U> C> G
12	A>C=G	A>U>C	A >U >G	C>G>U
...	...	...	...	...

Design rules obtained from experiments

- **Given**:
  - A set of siRNA with score of knockdown efficacy.
  - A number of experimental design rules
- **Find methods**:
  - To predict the score of any siRNA
  - To artificially creates siRNAs with high scores of knockdown efficacy.

HCV NS5A and IFN/RBV therapy

- NS5A is the protein most reported in interferon resistance (Gao, *Nature* 465, 2010).
- What is the remained enigmatic role of the domains II and III (Lemon, 2010)?
- V3 is a more accurate biomarker than the ISDR region (AlHefnawi, 2010)?
- Distinguish the responders and non-responders between subtypes 1a-c (448 aa) and 1b (447 aa).



The data and two methods for motif finding

- **Labeled data**:
  Los Alamos HCV database (134 RVR and 93 S5A non-SVR sequences.
- **Unlabeled data**:
  5000 NS5A sequences belonging to 6 genotypes, mostly in 1a-c and 1b taken from Genbank and Nagoya City University.

- **Given**
  - SVR and non-SVR samples
  - unlabeled samples
- **Find**
  1. All strong motifs of type DOOPS (discriminative one occurrence per sequence) for each class of SVR or non-SVR?
  2. Strong motifs of type DMOPS (discriminative multi occurrence per sequence) satisfied two complete and discriminant conditions.

Strong motifs in terms of coverage and discrimination ability

DOOPS motifs found

Table 1. Typical β-DOOPS motifs with high value of β in the V3 and ISDR regions.

V3 region							ISDR region						
Type	1ac			1b			Type	1ac			1b		
	SVR	nSVR	UL	SVR	nSVR	UL		SVR	nSVR	UL	SVR	nSVR	UL
# Seq.	14	73	1559	30	49	1159	# Seq.	14	73	1559	30	49	1159
SGC	2	66	678	0	0	1	ANH	1	61	649	0	0	3
GCP	2	63	592	0	0	1	CTA	1	61	673	4	3	47
PSG	2	57	455	0	0	5	TAN	1	61	653	0	0	3
APSG	1	57	392	2	0	2	AAN	0	25	87	0	0	0
NTTA	0	29	69	0	0	0	IAA	0	25	84	0	0	0
EPAS	0	9	300	0	0	0	LIA	0	25	85	0	0	0
							WNQ	0	25	0	0	0	0
PDC	7	0	28	0	0	0	ESES	0	24	6	0	0	0
DDT	5	0	204	0	0	2							
GDD	5	0	249	0	0	7	DAN	0	0	1	3	0	3
PPDC	7	0	9	0	0	0	LVD	0	0	0	3	0	1
SEPT	4	0	67	0	0	0	VDA	0	0	0	3	0	1
EPTP	4	0	35	0	0	0	CTTH	0	0	10	2	15	317
DQA	0	0	0	2	12	264							
QSA	0	0	0	1	11	277							

DOOPS motifs found

Table 2. Typical β-DOOPS motifs with high β in the domain I and domains II-III.

Domain I							Domain II, III						
Type	1ac			1b			Type	1ac			1b		
	SVR	nSVR	UL	SVR	nSVR	UL		SVR	nSVR	UL	SVR	nSVR	UL
# Seq.	14	73	1559	30	49	1159	# Seq.	14	73	1559	30	49	1159
EYP	1	58	846	0	0	1	AAN	0	25	86	0	0	0
LHE	1	58	848	0	0	39	LWN	0	25	0	0	0	0
RVD	1	18	20	0	0	3	NQE	0	25	0	0	0	0
CDF	0	20	4	0	0	0	WNQ	0	25	0	0	0	0
MWD	0	14	1	0	0	0	SES	0	24	6	0	0	7
CDF	0	20	4	0	0	0	SKV	0	24	6	0	0	15
APP	0	0	2	0	7	9	KNP	0	0	0	8	0	31
DYA	0	0	0	0	7	7	NPD	0	0	0	8	0	40
KNP	0	0	0	8	0	31	WKN	0	0	0	8	0	31
NPD	0	0	0	8	0	40	GEI	0	0	1	7	0	11
WKN	0	0	0	8	0	31							
SGT	11	59	1241	6	0	8							
WSG	11	59	1247	6	0	3							
DSH	0	0	42	6	0	71							
GDS	0	0	42	6	0	72							
SNM	0	0	2	0	7	213							

DMOPS motifs found

Fold 1		fold 2		fold 3	
SVR	nSVR	SVR	nSVR	SVR	nSVR
DAE	QA	AI	NM	AI	ND
AI	AEA	DI	AKA	DAN	GR
AAG	NM	AAC	RRRLA	TAA	NM
TRAL	GR	TRA	DT	HA	MA
AF	DK	RAC	IKA	FR	RS
CR	RF				DK
	DI				PNA
					EAT
					LGA
					AEA

Conclusion

- Computer science and computational science play an increasing important role in medicine.
- Might biomedical informatics be to public health in the 21st century what infectious diseases were to public health in the previous centuries.
- Learn and use more molecular biology in medicine by informatics.

THANKS

What makes proteomics important?

- There are more than 160,000 genes in each cell, only a handful of which actually determine that cell's structure.
- Many of the interesting things about a given cell's current state can be deduced from the type and structure of the proteins it expresses.
- Changes in, for example, tissue types, carbon sources, temperature, and stage in life of the cell can be observed in its proteins.