

Khoa học: Hai, ba hay bốn chân?

Câu chuyện về khoa học tính toán và khoa học với dữ liệu lớn

Hồ Tú Bảo



Vietnam Academy of Science and Technology
Viện Khoa học Tính toán và Điều khiển →
Viện Tin học → Viện Công nghệ Thông tin

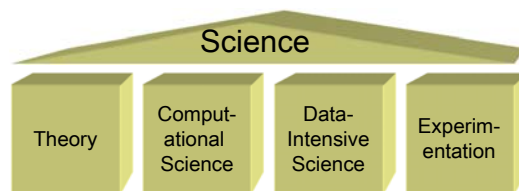


Steering Committee member
- PAKDD: Pacific Asia on Knowledge Discovery & Data Mining
- PRICAI: Pacific Rim Inter. Conf. on Artificial Intelligence
- ACML: Asia Conference on Machine Learning

Japan Advanced Institute of Science and Technology
School of Knowledge Science



Science: Two, three or four legs?



DOI:10.1145/1859204.1859206

Science Has Four Legs

CACM, Dec. 2010

DOI:10.1145/1810961.1810962

Science Has Only Two Legs

Science has been growing new legs of late. The traditional "legs" (or "pillars") of the scientific method were *theory* and *experimentation*. That was then. In 2005, for example, the U.S.

CACM, Sep. 2010



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

1. Cảnh báo vị trí dẫn đầu và nhận định sức cạnh tranh của Mỹ sẽ phụ thuộc vào sự phát triển khoa học tính toán.
2. Cấu trúc của nghiên cứu và giáo dục cho thế kỷ 21
3. Lộ trình nhiều thập kỷ của khoa học tính toán.
4. Hạ tầng cơ sở bền vững cho sáng tạo và cạnh tranh.
5. Thách thức của nghiên cứu và phát triển.
6. Một số lĩnh vực của khoa học tính toán



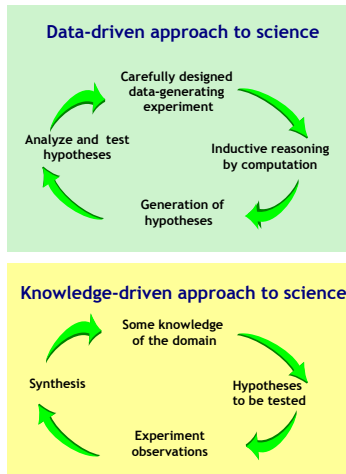
PITAC: President's Information Technology Advisory Committee.

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

Data-intensive science: a shift in science

Khoa học với dữ liệu lớn dùng tiếp cận “data driven”, nhằm tìm thông tin và tri thức từ dữ liệu.

Tiếp cận truyền thống “knowledge driven”, nhằm kiểm nghiệm các giả thuyết khoa học hình thành từ tri thức.



Book: The Fourth Paradigm, 2009 & Newman et al., CACM 2003

5

Content



- Khoa học tính toán & Khoa học với dữ liệu lớn
 - Mô hình và mô phỏng
 - GPGPU và siêu máy tính
 - Dữ liệu lớn, máy học và khai phá dữ liệu
 - Ngôn ngữ, tài chính và sinh học tính toán
- Nghiên cứu trong biophysics và biomedicine
 - Cấu trúc và động học của phân tử nước trong các hệ sinh học
 - Khai phá dữ liệu y-sinh trong nghiên cứu bệnh viêm gan
- Trao đổi về nghiên cứu



6

Computer science and computational science

- Information Technology: Cùng nghĩa với Tin học (informatics) và Khoa học máy tính (Computer Science): khoa học và công nghệ về xử lý tự động thông tin trên máy tính.
- Xử lý thông tin là quá trình biến đổi dữ liệu từ dạng này sang dạng khác để thu được thông tin và tri thức mới.
- Computational science: Khoa học về xây dựng và thực hiện các mô hình tính toán để giải bài toán trong các khoa học khác.
- Thí dụ từ PITAC report
 - Social sciences
 - Physical sciences
 - National security
 - Geosciences
 - Engineering/Manufacturing
 - Biological sciences and Medicine

Difference between Science, Technology, and Engineering?

7

Computational Science

“Là một lĩnh vực liên ngành, nhằm dùng các khả năng tính toán tiên tiến để hiểu và giải quyết các bài toán phức tạp”.

Khoa học tính toán hòa trộn 3 thành phần:

- Các thuật toán và phần mềm về mô hình hóa và mô phỏng để giải các bài toán về khoa học, kỹ nghệ và xã hội.
- Khoa học máy tính và thông tin nhằm phát triển và tối ưu các hệ thống tiên tiến về phần cứng, phần mềm, mạng máy tính và quản trị dữ liệu để giải các vấn đề cần đến tính toán.
- Hạ tầng cơ sở tính toán cần cho việc giải các bài toán khoa học và kỹ thuật cũng như để phát triển khoa học máy tính và thông tin.

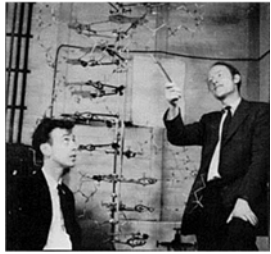


PITAC's report, 2005

8

Model and Modeling

- Model: Mô tả khái quát của một thực thể (Simplified description or abstraction of a reality).
- Modeling: Quá trình tạo ra một mô hình.



DNA model figured out in 1953 by Watson and Crick

- Mô hình giao thông tại Hà Nội?
- Mô hình thị trường và giá cả?
- Mô hình diễn biến một dịch bệnh?
- Mô hình một xã hội điện tử?
- Mô hình tác chiến hợp đồng binh chủng trong một chiến dịch?

Grande challenges in modeling?

Simulation

- Mô phỏng: Là việc tạo ra như thật trên máy tính các thực thể sao cho có thể thấy chúng xảy ra thế nào.
- Internet Simulator StarBed: Hệ mô phỏng có simulator lớn nhất thế giới, có thể mô phỏng 8,000 nodes



StarBed: Network simulation, a JAIST project

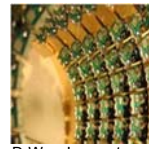


Competition on high performance computing

- Japanese project on the Fifth Generation of Computer Systems (1982-1992)
- Quantum computers
- PC clusters
- GPGPU
- Supercomputers



JAIST's PIM (Parallel and Inference Machine)



D-Wave's quantum computer ready for latest demo



Our GPGPU computer



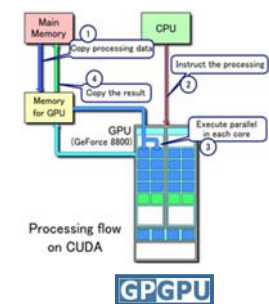
JAIST's CRAY XT5



Our PC clusters

GPGPU: General purpose GPU

- GPU được phát triển thành những bộ xử lý hết sức linh hoạt và mạnh về khả năng lập trình, độ chính xác, hiệu quả. GPGPU = mở rộng GPU cho các nhiệm vụ tính toán khác.
- Khó khăn của lập trình GPGPU được khắc phục bởi 'C for CUDA' (Compute Unified Device Architecture, 'koo-duh') của NVIDIA (available: Python, Fortran, Java, MATLAB).
- Giá rất rẻ (> 100M CUDA-enabled GPUs sold)



Tesla C20, 520 ~ 630 GFLOPS
2,495 USD

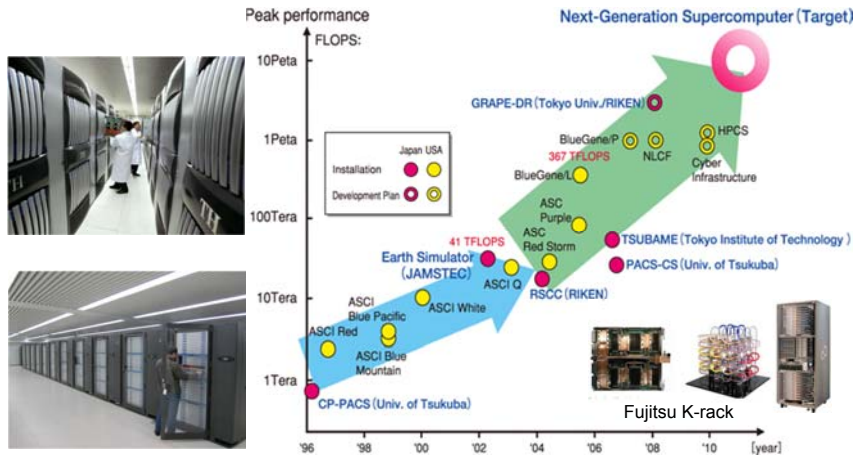
Tesla S20
2.1 ~ 2.5TFLOPS
12,995 USD

Cray XT-3, 2.2 TFLOPS

<http://www.nvidia.com/content/global/global.php>

Next generation supercomputer projects

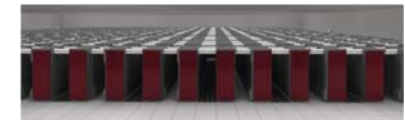
Japan national key project (2007-2012)



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

Next generation supercomputer projects

- 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>
 (23 Nov. 2010)

Next generation supercomputer projects

Do chưa phát triển được các ứng dụng, siêu máy tính sẽ không tạo được nhiều tác động vào nền kinh tế và cuộc sống thường nhật của chúng tôi. Thậm chí nếu phát triển được các ứng dụng ở Trung Quốc, chúng tôi vẫn đi sau Mỹ 5 năm vì không biết cách sử dụng những ứng dụng đó.

CNN, Nie Hua, Dawning Info. Industry (about Tianhe-1A, Nebulae)

Kennichi Miura, DEISA Symposium, 5.2007

How much data is there?

Large Hadron Collider, (PetaBytes/day)

Human Genomics = 7000 PetaBytes
1GB / person

1 human brain at the micron level = 1 PetaByte

1 book = 1 MegaByte

Family photo = 586 KiloBytes

Kilo	10^3
Mega	10^6
Giga	10^9
Tera	10^{12}
Peta	10^{15}
Exa	10^{18}

The LIBRARY of CONGRESS
Printed materials in the Library of Congress = 10 TeraBytes

200 of London's Traffic Cams (8TB/day)

All worldwide information in one year = 2 ExaBytes

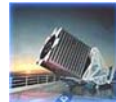
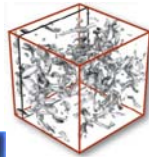
Adapted from Berman, Director of SDSC (San Diego Supercomputer Center)

Science paradigms

- Hàng nghìn năm trước: khoa học là **thực nghiệm**
Mô tả các hiện tượng thiên nhiên
- Vài trăm năm vừa qua: thêm nhánh **lý thuyết**
Dùng các mô hình, các khái quát hóa
- Vài thập kỷ vừa qua: thêm nhánh **tính toán**
Mô phỏng các hiện tượng phức tạp
- Ngày nay: **Khai thác dữ liệu** (eScience)
Hợp nhất lý thuyết, thực nghiệm và mô phỏng
 - Dữ liệu từ đo đạc bằng máy hoặc mô phỏng
 - Được xử lý bởi các phần mềm
 - Thông tin và tri thức được chứa trong máy tính
 - Nhà khoa học phân tích cơ sở/tập dữ liệu với công cụ quản trị dữ liệu và thống kê.



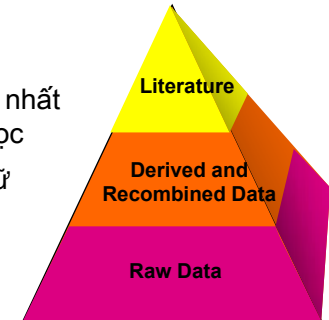
$$\left(\frac{d}{a}\right)^2 = \frac{4\pi G\rho}{3} - K\frac{c^2}{a^2}$$



The Four Paradigm: Data-Intensive Scientific Discovery, 2009

All Scientific Data Online (Jim Gray)

- Nhà khoa học được dùng dữ liệu từ nhiều lĩnh vực liên quan
- Internet là môi trường cho phép hợp nhất mọi nguồn tài liệu và dữ liệu khoa học
- Nhà khoa học sẽ đi từ tài liệu đến dữ liệu và tính toán rồi trở về với tài liệu
- Thông tin khoa học của mỗi người sẽ đến được bất kỳ ai khác, bất kỳ nơi nào khác
- Tốc độ của thông tin khoa học tăng lên rất nhiều do vậy năng xuất khoa học tăng lên bội phần

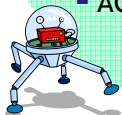


"You may say I am a dreamer but I am not the only one"

18

Machine learning and data mining

- Machine learning**
To build computer systems that learn as well as human does.
 - ICML since 1982 (23th ICML in 2006), ECML since 1989.
 - ECML/PKDD since 2001.
 - ACML starts Nov. 2009.
- Data mining**
To find new and useful knowledge from large datasets.
 - ACM SIGKDD since 1995, PKDD and PAKDD since 1997
 - IEEE ICDM and SIAM DM since 2000, etc.

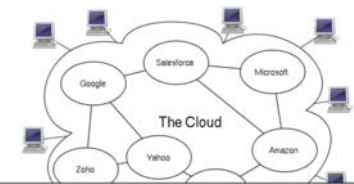


My group on Machine Learning & Data Mining: 7 PhD students, 5 master students

19

Some emerging trends in computing

- Services computing: Nhằm mô hình, sáng tạo, tác nghiệp, và quản lý các dịch vụ kinh doanh.
- Cloud computing: Liên quan những ứng dụng như dịch vụ trên như các phần cứng ở các trung tâm cho các dịch vụ
- FOSS: Free and Software



Contribution to Japan GDP in 2007 of "knowledge industry" is 69.5%

	Contribution to GDP (billion yen)	%
Agriculture and Forestry/Fishers Industry	7437	1.4
Mining Industry	435	0.08
Manufacturing Industry	107766	20.5
Construction Business	31849	6.0
Electronic Power/ Gas / Water Service	11565	2.2
Commerce	68234	12.9
Finance and Insurance	35207	6.7
Real Estate Industry	60465	11.5
Traffic/ Information, Communication	33524	6.3
Public/Governmental service	47306	9.0
Service	110695	21.1
Non profit service producers	10710	2.0
Total GDP	525191	100%

The best way to predict the future is to invent it (Alan Kay)

20

Computational finance

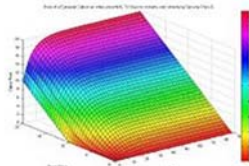
Lĩnh vực liên ngành của trí tuệ tính toán, tài chính toán học, phương pháp số và mô phỏng để tạo các quyết định thương mại (trading), phòng ngừa thất thiệt (hedging) và đầu tư (investment), cũng như làm dễ dàng quản lý rủi ro của các quyết định trên.

$$C = SN(d_1) - Ke^{-r(T-t)}N(d_2)$$

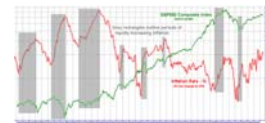
where

$$d_1 = \frac{\ln(S/K) + (r + \frac{\sigma^2}{2})(T-t)}{\sigma\sqrt{T-t}}$$

$$d_2 = d_1 - \sigma\sqrt{T-t}$$

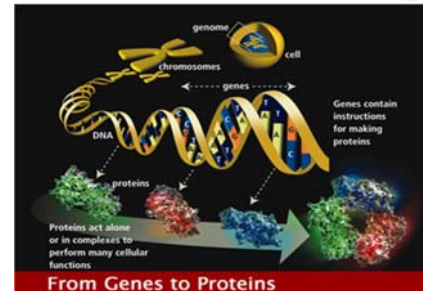
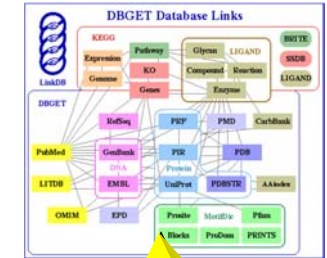
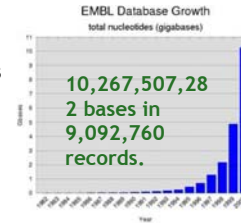


Black-Scholes European Call Option Pricing Surface

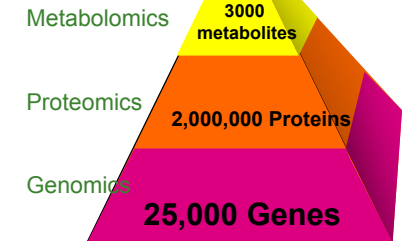


Computational biology

One of very few cases of human activities growing faster than Moore's law.



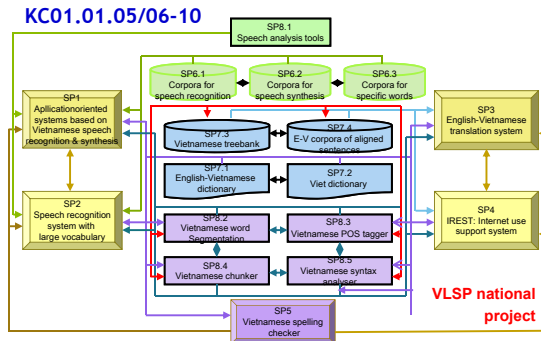
Biology in the 21st Century



Computational linguistics



Voice search
[...downloadpapers/Video Clips/Google Mobile App for iPhone with Voice Search.mp4](#)



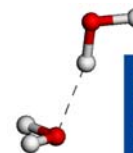
Google Translation



<http://vlsp.vietlip.org:8080/demo/?page=about>

What is the structure of water?

One of the 100 outstanding unsolved problems in science



Is superfluidity possible in a solid? If so, how?
 Despite hints in solid helium, nobody is sure whether a crystalline material can flow without resistance. If new types of experiments show that such outlandish behavior is possible, theorists would have to explain how.

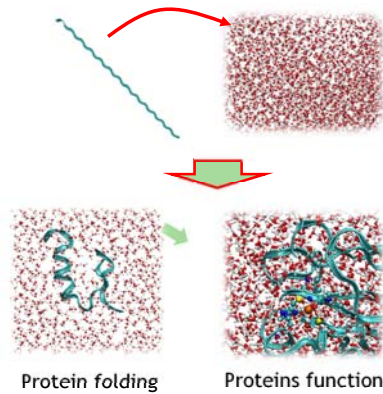
JUPITER IMAGES

What is the nature of the glassy state?
 Molecules in a glass are arranged much like those in liquids but are more tightly packed. Where and why does liquid end and glass begin?

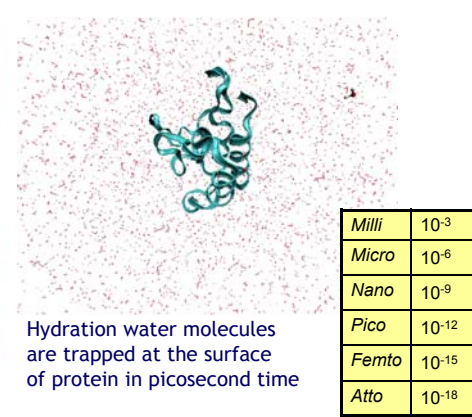
What is the structure of water?
 Researchers continue to tussle over how many bonds each H₂O molecule makes with its nearest neighbors.

Water in biological systems

Water and Proteins

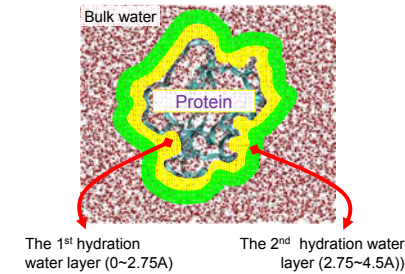


Hydration Water



Views on bulk water and hydration water

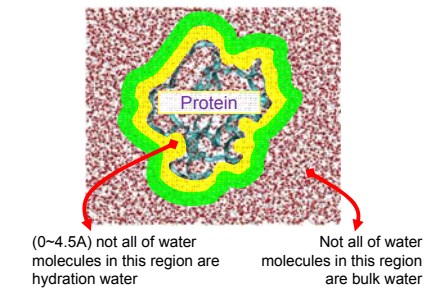
The common used definition



Strongly depend on water relative motions and their relative distance to the protein surface, with fixed hydration shell.

Chen et al., Physical Chemistry B, 112, 2008.

Our view by dynamic behaviors

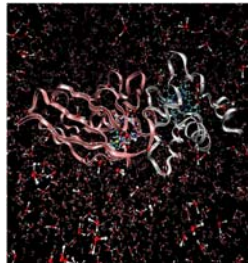


Grouping water molecules into two classes using their dynamical behaviors.

New direction: Simulation-based data mining

..\\ResearchCollaboration\WaterProperties\J2M_hbond.mpg

SIMULATION



DATA

Atom 1 ~6Tb

Atom 2

t_1	x_1	y_1	z_1
t_2	x_2	y_2	z_2
\vdots	\vdots	\vdots	\vdots
t_n	x_n	y_n	z_n

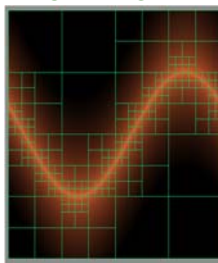
Atom N

t_1	x_1	y_1	z_1
t_2	x_2	y_2	z_2
\vdots	\vdots	\vdots	\vdots
t_n	x_n	y_n	z_n

$n \sim 2 \times 10^5$

$N \sim 3 \times 10^4$

KNOWLEDGE

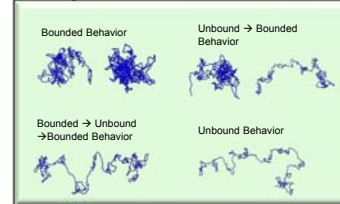


M. Better et al., IBM (2007): market research application
 M.K. Painter et al., (2006): aircraft engine fleet management

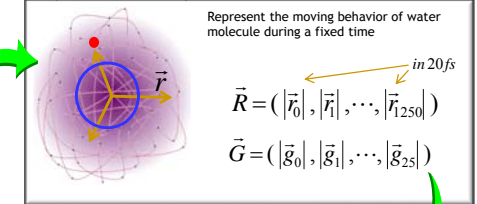
Dam Hieu Chi, Ho Tu Bao et al. (2010): Toward a simulation-based approach to data mining in scientific domains.

Discovery of water structure and dynamics

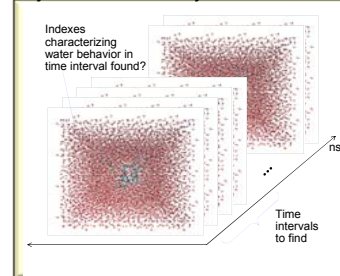
Moving behavior



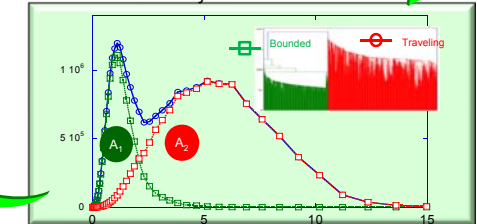
Representation space



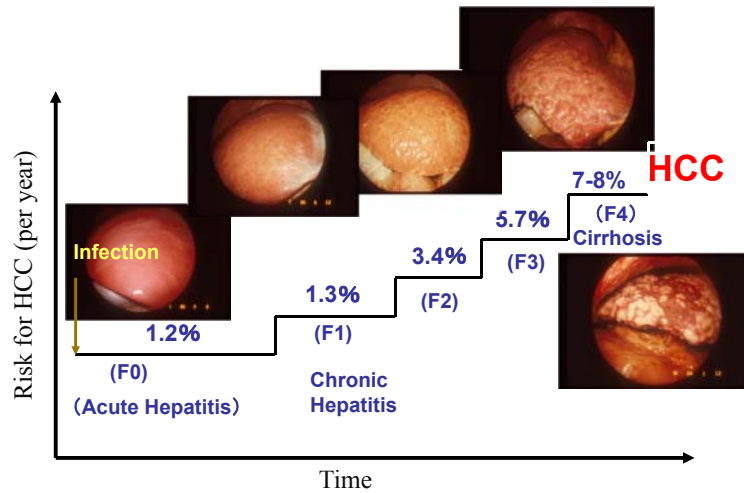
Dynamics discovery



Structure discovery

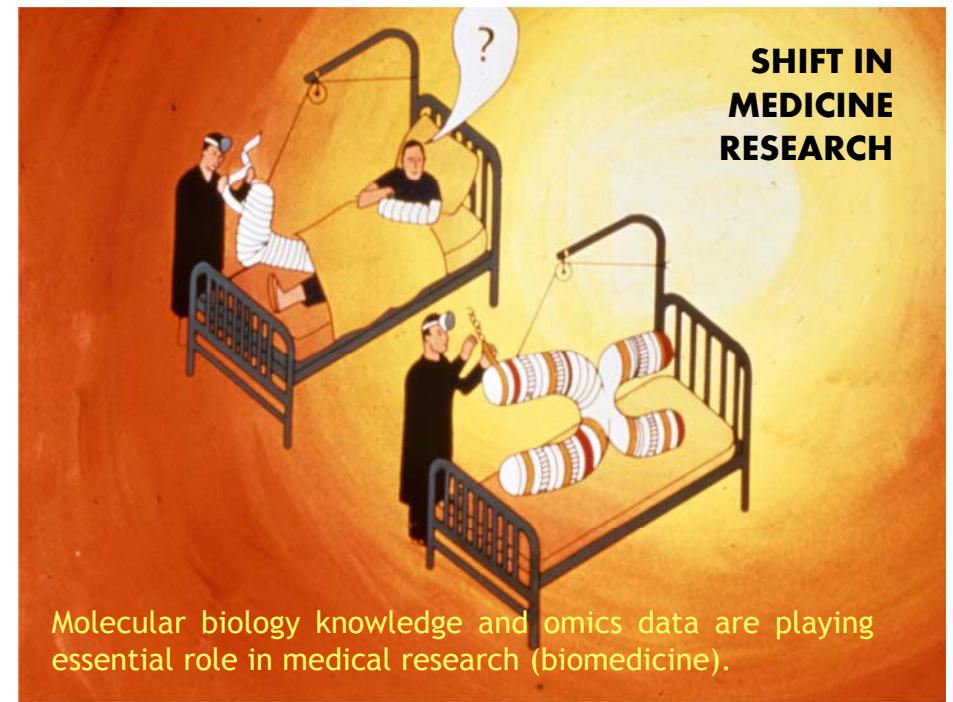


Fibrosis and HCC in HCV



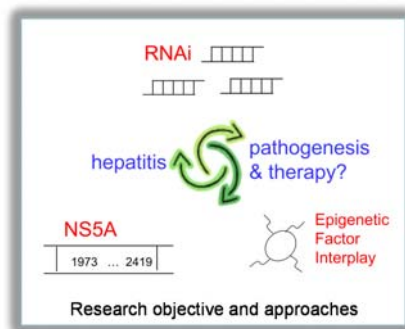
Percentage: rate of HCC occurring from F0, F1, F2, ... staging of fibrosis.

29



Limitations in computational biomedicine

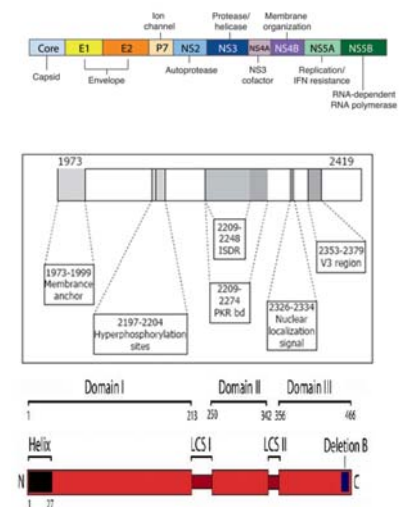
- Use small data sets taken at a hospital. Some with high-throughput profiling data but no link to hospital data.
- Use basic statistics but not advanced machine learning and data mining.
- Only 50% HCV are cleaned by IFN/RBV.
- Aim to provide that basis that multiple therapeutic strategy can increase the success rate of hepatitis treatment.



31

HCV NS5A and IFN/RBV therapy

- NS5A inhibitor is a hot topic in study of virus C genome and their drug resistance mechanisms.
- **Problem:** Molecular mechanisms of NS5A resistance to IFN/RBV therapy?
- NS5A inhibits IFN α activity via its interaction with IFN α cellular antiviral pathways and the mutations in NS5A resist IFN α therapy.
- Some recent questions
 - What is the enigmatic role of the domains II and III (Lemon, 2010)?
 - V3 is a more accurate biomarker than ISDR region (AlHefnawi, 2010)?

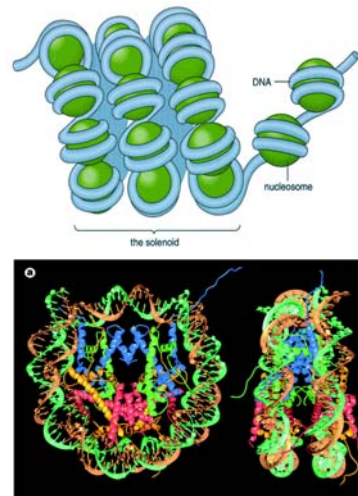


Gao, *Nature* 465, 2010

32

Epigenetics and hepatitis therapy

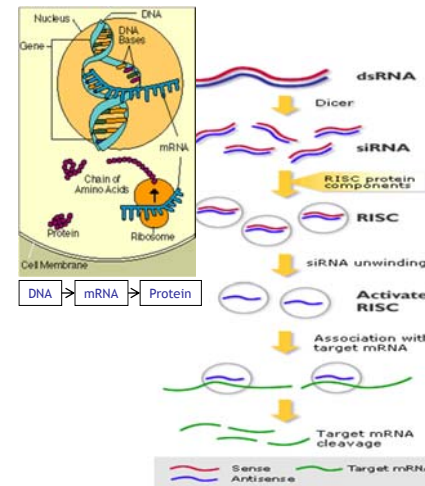
- Heritable changes not only in primary nucleotide sequence
→ Epigenetics
- Two major epigenetic mechanisms:
 - DNA methylation (addition of a methyl group to 2 bases of DNA)
 - histone modification (mono-, di-, or tri-modifications of protein after translation).
- Problem:** Molecular mechanisms of epigenetic modifications and their crosstalk and their relations to the development of the diseases, esp. liver cancer.



Luger et al. Nature, (1997)

33

RNA interference (RNAi) and hepatitis



- RNAi is post-transcriptional gene silencing (PTGS) mechanism.
- RNAi (siRNA and miRNA) target to HBV and HCV genes to inhibit their replication or host genes required for their replication.
- Chemically synthesized siRNAs can mimic the native siRNAs produced by RNAi but having different ability.
- Problem:** Selection of potent siRNAs for silencing hepatitis viruses?

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

34

Table 5. The 0.95-discriminant patterns found in ISDR region (the last is 0.88 discriminant pattern).

Type	Iac R	Iac NR	Iac UL	Ib R	Ib NR	Ib UL
# Seq.	14	73	1559	30	49	1159
ANH	1	63	649	0	0	3
CTA	1	63	673	4	3	47
TAN	1	63	653	0	0	3
AAN	0	25	87	0	0	0
IAA	0	25	84	0	0	0
LIA	0	25	85	0	0	0
LWN	0	25	0	0	0	0
NQE	0	25	0	0	0	0
WNQ	0	25	0	0	0	0
ESES	0	24	6	0	0	0
DAN	0	0	1	3	0	3
LVD	0	0	0	3	0	1
VDA	0	0	0	3	0	1
KATCTH	0	0	10	2	15	317

Table 7. The 0.95-discriminant patterns found in the domain II and III.

Type	Iac R	Iac NR	Iac UL	Ib R	Ib NR	Ib UL
# Seq.	14	73	1559	30	49	1159
VEI	0	27	272	0	1	2
AAN	0	25	86	0	0	0
LWN	0	25	0	0	0	0
NQE	0	25	0	0	0	0
WNQ	0	25	0	0	0	0
SES	0	24	6	0	0	7
SKV	0	24	6	0	0	15
KNP	0	0	0	8	0	31
NPD	0	0	0	8	0	40
WKN	0	0	0	8	0	31
GEI	0	0	1	7	0	11

Some results on NS5A

Table 6. The 0.95-discriminant patterns found in the domain I.

Type	Iac R	Iac NR	Iac UL	Ib R	Ib NR	Ib UL
# Seq.	14	73	1559	30	49	1159
EYP	1	58	846	0	0	1
HEY	1	58	845	0	0	38
LHE	1	58	848	0	0	39
RVD	1	58	70	0	0	3



35

Conclusion

- Dù xem là có ba hay bốn chân hay không, khoa học của thế kỷ 21 sẽ phát triển nhờ khoa học tính toán và khoa học với dữ liệu lớn.
- Biết** những xu thế và tiến bộ của CNTT, biết linh hoạt và thay đổi để tìm thấy đường của mình.
- Định** được những vấn đề có ý nghĩa khoa học, giá trị thực tiễn cao và có tính khả thi ở Việt Nam.



Thanks

36