Hanfei Bao: The Theory of Biomedical Knowledge Integration(Ⅰ), Chinese Journal of Medical Treatment, Vol(2):No(13), 1-7, 2003

The Theory of Biomedical Knowledge Integration(Ⅰ)

Hanfei Bao
Lab of Informationization and Standardization of TCM
Shanghai Univ. of Traditional Chinese Medicine
Shanghai, China, bhf2002@online.sh.cn

The innumerable biological and medical data, information and knowledge have been found in the scientific activities in biomedical fields. All of them are coming from the identical "mystical lands", ie the lively, energic and sophisticatedly structured and ordered organisms. Therefore these data, information and knowledge are "brothers and sisters", should not be the strangers to and isolated from one another. How can they communicate, link and be organized or united like they originally did in the organisms? When can they come back to their lovely homelands hand in hand? How possible can we fully use them to reconstruct or integrate logically or physically the models of the organisms? Biomedical Informatics[1, 2] now perhaps has been facing these questions and will be duty-bound to undertake the new important mission, the large scale integration of biomedical data, information and knowledge. Let's first call the new efforts in this direction Biomedical Knowledge Integration (BMKI). Albeit perhaps some body would say it might be a work in the very remote future and may be he is quite right, it is never too early for people to do some exploratory investigation. In this series of papers the term "knowledge" is defined as the joint name of three widely used and polysemous concepts data, information and knowledge.

Both the significance and difficulty of this new mission for Biomedical Informatics might be beyond imagination and comparison. It needs much more efforts of Biomedical Informatics, Biology and Medicine in many fields.

Ⅰ. Organisms——The Natural Integration Master

It is the scientific activities of human being that make our organisms being continuously divided into more macro-micro levels, from quantum biology, molecular biology, biophysics, biochemistry, through cytology, histology, anatomy, physiology, immunology, genetics, to clinical sciences. Whereas the organisms themselves are doing contrarily. They arrange thousands of elements and mechanisms orderly, hierarchically, and jointly in the very limited spaces in the physical bodies. This magical work of nature is some kind of the physical integration and MBKI would be contrastedly an artificial intelligent integration. So I would say that MBKI, which is trying to integrate or link all of the biomedical knowledge we have acquired, obeys

the will of nature.

We can't help admiring so much the fascinating abilities of the organisms to carry out the task of natural or physical integration. In order to show the readers a bit of integration talent of them, the author has summarized the dynamical circles, which perform orderly and jointly to complete a whole blood pump, from molecular level to integral level of heart[3,4]. These circles are drived by the histological structures and physiological mechanisms, such as ion channel, ion pump, membrane electronic potential, action electronic potential, endoplasmic reticulum and calcium ion, ATP enzyme, myofilament protein(thick myofilament and thin myofilament), neurotransmitter-receptor, synapse, myocardial cell, fluid mechanics, stop-cock principle, etc.(see Fig. 1) A normal blood pump relies on the precisely and serially linking up of those circles and any phase mismatching between the circles might lead to malfunction of the heartbeat.

What a perfect and miraculous physical integration!

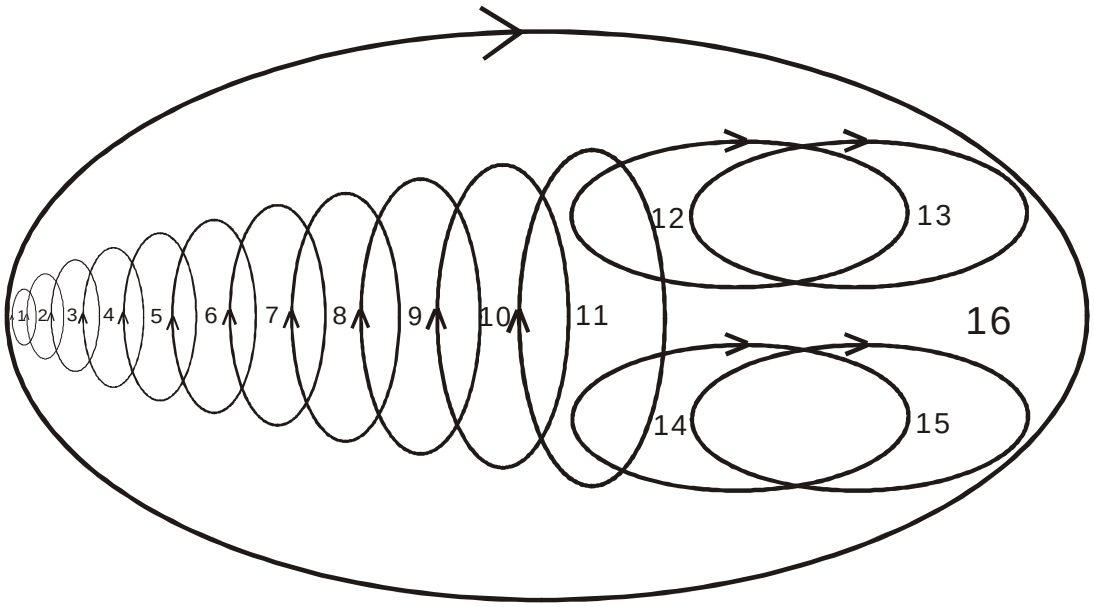


Fig.1 The linking topology of the heart blood pump circles from molecular level to organ-integral level: 1.membrane potential circle; 2.circle of Ca^{++} between sarcoplasm and sarcoplasmic reticulum(by Ca^{++} pump); 3.troponin and Ca^{++} combining-releasing cirlce; 4. troponin molecular configuration change circle; 5.tropomyosin molecular configuration change circle; 6.cross bridge and actin combining-bending-releasing circle; 7.thick myofilament and thin myofilament sliding circle; 8.sarcomere contraction-expansion circle; 9.muscle fiber contraction-expansion circle; 10. muscle contraction-expansion circle; 11.blood pressure circle in ventricle; 12.open-close circle of atrio-ventricular valve; 13. blood pressure circle in atrium; 14. open-close circle of aortic valve; 15. blood pressure circle in aorta; 16.blood pump of heart.

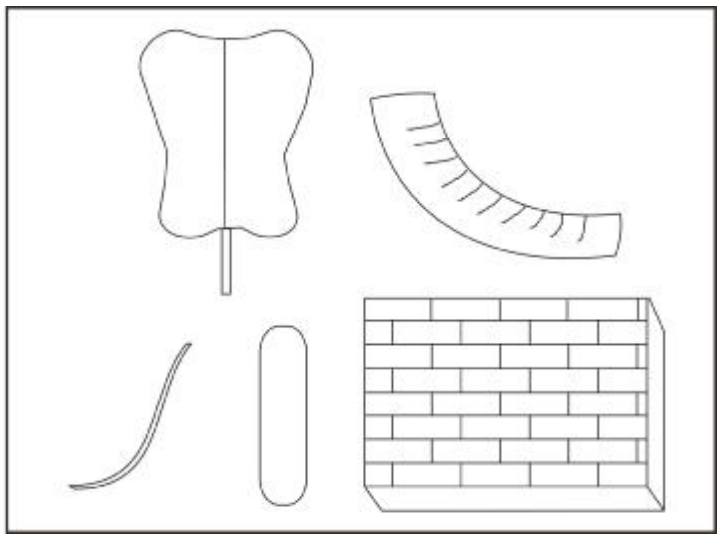
Ⅱ. The Revised Story of Blind Men and Elephant

A well-known story of Blind Men and Elephant told that one day five blind men was feeling an elephant with the hands and trying to say what the animal is like. One was touching the elephant's nose and declared "Elephant is like pipe". Another blind man, who was handling its tail, thought it was similar to rope. And the other three considered the animal was something like "post", "wall" or "fan", when they felt the elephant's leg, trunk and ear, respectively(see Fig.2 A). This story has been historically a funny material to tell people how nonsense to use only the local knowledge to explain the whole.

Whereas the things have been changed. The blind men have realized that "knowledge integration" is important and they started to find a way to link their knowledge together and have at last made great progress. The result from the integration is: "A wall is the centre of the elephant, a pipe and two fans are attached in front, a rope linked to the rear and four posts fasten to the lower part of the wall"(see Fig.2 B). Thus through knowledge integration, the understanding of our blind group has got much closer to the true elephant, although the information own by each person has not changed much.

Through this invented story, the author try to explain how significant of BMKI will be for the whole understanding of the biologic systems and human body. We should not only and forever fix our eyes on the more and more detailed details of the life systems. We should also pay our energies on examining those existing numerous, heterogenous and isolated (under physical semantic sense) data, information and knowledge we have got in biomedical fields, to see what we can do in the aspect of reassembling them integrally, what fundamental problems we will meet on that way, what an extent to which we can reach in this direction, where the black abysses cross which people can hardly jump are waiting for us and what points we can begin with. We should always bear in mind that these data, information and knowledge are from the organisms arising from the identical bio-evolutional tree. No matter the purposes, perspectives and methods, under which those data , information and knowledge were obtained, are so different.

A



B

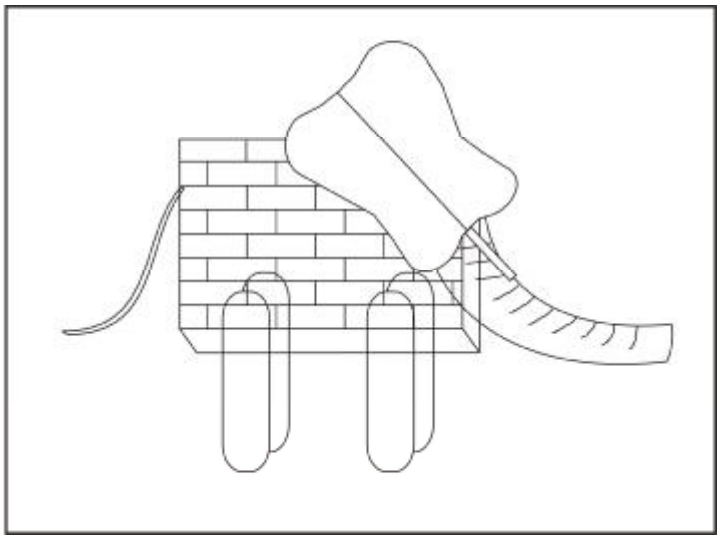


Fig. 2 A What the blind men imagined about the elephant before they doing knowledge integration; B the elephant imagined by them afterwards.

Ⅲ. The Four-Incompleteness-Principle in Cognition Processes

Human being acquires the knowledge or understanding of the world through the cognition processes. Thus there are inherent relations between the knowledge or knowledge integration and cognition sciences. There are at least four key aspects in the cognition processes: (1)the cognition object, ie the world; (2) the cognition context, condition or tools; (3)the knowledge expression; (4)the awareness of the knowledge in our brain. Unfortunately none of the four aspects could be complete for the cognition processes. This property of the cognition processes of human is called in this paper the Four-Incompleteness-Principle.

(1) The incompleteness of our knowledge about the world

This incompleteness is because that the world, ie the cognition object, always tells us parts of its story. The world, including human being himself and the cognition process itself, is a "shy girl". She always "hides her half face behind Pipa"(a kind of ancient Chinese music instrument), keeping the other half of her face out of our sight. Forever, the world can be partially observed and understood by us. That is the reason why the scientists, generation by generation, have dedicated themselves to its secrets.

(2) The incompleteness of awareness of the background of knowledge acquirement

The background or context means here the conditions or methods through which we have obtained the knowledge. As we known, many units of knowledge have been achieved by

observation or measurement under very special or particular clinical or experimental conditions and our brains usually logically or even at will omit them and use those knowledges under the general sense.

(3) The incompleteness of the knowledge expression

That is because the media usually express the knowledge also partially. Suppose there is a very traditional Chinese who knows nothing about the cultural backgrounds of Europe, he might get many troubles when he reads a translation of an Europe novel. Because according to the generalized economic principle which plays a role everywhere, the author of the novel didn't need to write down everything especially those "understandable without saying". Otherwise his work would not be considered as succinct. Those left out and "understandable without saying" portions of the story are the potential parts of things for the European, but would be problematic for this traditional Chinese. That is to say there are two attributes, ie the manifestation and potentiality, in knowledge expression. That causes the incompleteness of the knowledge expression. In the scientific data, information and knowledge, in fact, a large quantity of special contexts is left out, because they are "understandable without saying" for the domain exerts, but usually not for the experts in other domains, especially not for computers. Here is another most simple example of physiological knowledge. A textbook of physiology says "the means of the adult heart rate and the duration of adult heart cycle are 75 times/m and 0.8 sec respectively . " It implies the proposition is based on the physical context: general adult (no patient, no athlete etc), during rest(not exercise, not being excited), etc.

Any knowledge has its (obvious or hidden) context or condition which guarantees its validity. Famous Newton's three laws are valid in world of general size and general speed, but invalid in both quantum and light-speed worlds. Additionally, we could not talk about what the world is like beyond the scientific levels, methods and conditions at that time. As the cognition subject, our brain gets the data from the original world always through certain means or tools of observation or measurement, which could be auditory, optic and contact organs or receptors of human and the electronic, magnetic, electro-magnetic, chemical, etc advices as well. Those means or tools in turns determine the natures of data observed or measured. And, in fact, the means or tools themselves are the components of the generalized context of knowledge.

（4）The incompleteness of awareness of the knowledge in brain

Once a baby came into the world, it starts the cognition processes immediately in a way which is utterly different from those of other kinds of animal. So the special, basic abilities and patterns of cognition exist congenitally.

Similar to that the world keeps its portions from our sight stated above, the knowledge we actually own hides its portions, especially those gained congenitally, from our feeling, that is some of our knowledge exists in our brain subconsciously. We are not ware of their existence in our mind. It means that as a matter of fact we know more than what we can tell out clearly about our knowledge. That makes many differences between the knowledge structures of domain expert and expert system. To make it clear that how much inborn and acquired knowledge has been stored in our brain is by no means an easy job. We can hardly thoroughly list all the pieces of knowledge which take part in our mind work. There always the substantial parts of it are subconscious. In one word, the knowledge of which we are clearly aware of is never complete. That is the fourth incompleteness of the cognition processes the author would like to point out here. That makes many differences between the knowledge structures of domain expert and expert system.

Fig.3 shows two attributes of manifestation and potentiality at the significant links of people's cognition processes. It to a certain extent indicates the high complexities in biomedical artificial intelligence and biomedical knowledge integrations.

In the case the cognition subject is man, sometimes the four incompletenesses mentioned above may be not a problem at all, whereas for the computer in AL or MBKI, the situation might be much more serious and things could become much more complicated.

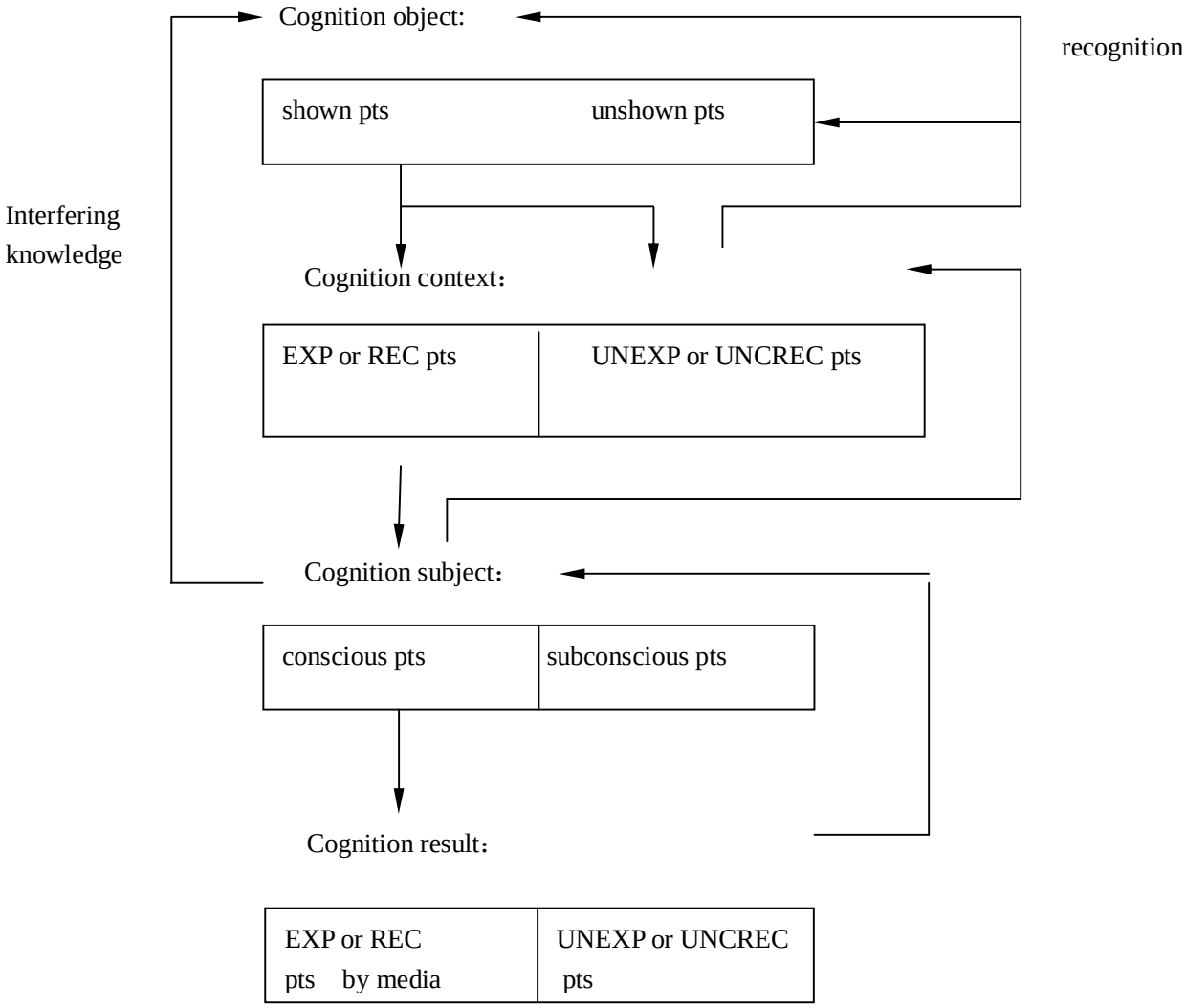


Fig.3 A kind of topologic figure showing the attributes of manifestation and potentiality in the human cognition aspects: the cognition object indicates the real world, cognition context means the particular circumstance, methods whereby the knowledge is obtained, cognition subject and cognition result represent human brain and knowledge-recording media, respectively. EXP or REC pts mean the expressed or recorded parts of knowledge and UNEXP or UNREC pts, the knowledge parts which are not expressed or recorded.

Ⅳ. The Basic Concepts of the Theory of Biomedical Knowledge Integration

The word "integration" is a polysemant. On the different occasions, it might arbitrarily be equivalent of such as "synthesis", "being concerned together", "being linked together", etc. Sometimes it means things physical but in other cases it means conceptually. In this article series, however, the author try to discuss the biomedical knowledge integration under rather rigorous physical or logical senses. It is expected to do some fundamental investigations and explorations on the biomedical knowledge integration problem. The last goal of BMKI is to help making the biomedical models as seamless as possible and being continuously close to original organism.

The all parts of a thing is called an whole, which has both philosophical and operational meanings. From the operational rather than philosophical perspective, an whole, contrasted to its parts, depends on our view angle or the purpose of the operations. For examples, in different cases, a person, family, company, country and the earth might be taken as either an whole or a part. Thus the concept whole is a relative one.

If two things are not independent of each other, then we say there is/are relation(s) between them. If the relations exist between two things, then we call them binary relations. The binary relations may be considered consisting of typically four essential parts: the generalized relational operator, subject set(τ), object set(ο) and obvious or obscure condition set(γ). Here set means a group things or elements playing jointly a common role in the relation, with or without structure, and an element is a smallest semantic unit. The generalized relational operator is the logical or physical action factor of a relation and reflects its nature, in most cases it corresponds to the verb between the subject and object in a sentence in Linguistics. Therefore the relational operator is the logical or physical "force" which causes the relation to be realized. The type of the logical relational operator or action factor sometimes reflects the view-angle, thereby is with somewhat arbitrary. The things which take the action make up the subject set and those receive the action are called the object set. So called the condition set is the indirect part of the relation, the role of which is to guarantee the bring about of the relation.

The Human-Computer Shareable Language(HCSL) [5-7] developed by the author with C language for the Integrateble and Relationized Medical Book(IREMB), where some examples for logical integration of medical knowledge were presented, has summarized ten generalized relations of medical knowledge according to the logical and especially the operational characters. The ten basic relations are pan-create(CREAT), pan-increase(INCRS), pan-decrease(DECRS), pan-contain(CONTN), pan-company(COMPN), pan-transform(TRANS), pan-equal(EQUAL), pan-order(PORDR), pan-pass(PASTO), pan-zero-relation(PNULL).

The relations in the semantic network of UMLS(Unified Medical Language System) [8] directed by NLM are divided into ①the subjective types: eg "is_a", "conceptually_related_to" and ② the objective types: eg "physically_related_to", "spatially_related_to", "functionally_related_to", "temporally_related_to". More than forty particular relations of

semantic network of UMLS are published on Internet and they might be reduced into less definitely defined relations based on their properties as the potential AI or BMKI operators. In UMLS the elements which may compose the subject, object and condition sets of the relations are split into entities(rather "hard" things) and events(rather "soft" things). Both entity and event are subdivided further into the more detailed items based on their qualities, such as subjective(mind-producing, conceptual, etc)-objective(real, physical, etc), normal-abnormal, natural-artificial, qualitative-quantitative, structural-functional, macro-micro, embryonic or full-developed, etc. The certainty and operational characteristics of those relational operators and elements in BMKI might be much different from each other, one of the reasons for that is the generalized degrees or the granularities of them being different.

"Relation integration" is almost another way saying "Knowledge integration", meaning the joint, systematical realization of a set of relations in logical or even physical way. Namely relation integration is a kind of multi-to-one transform or operation. Some basic principles of relation integration will be discussed in next article.

The operations of relation integration may be mathematical, general logical, biomedical logical, physical, etc. The logical operations follow the laws of thinking sciences which are interested in the behaviour of our thoughts whereas the physical operations mean here the natural or biological ways by which the living system is running on. Here is an example for logical relation integration. If we have relations A,B,C,D,……∈ F, then we have an integration (A, τ, o, γ) ∪ (B, τ, o, γ) ∪ (C, τ, o, γ) ∪ (D, τ, o, γ) ∪……∈ (F, τ, o, γ). The protein-protein interaction network[9] is an example of this type of relation integration. In molecular biology, we use two-hybrid approach to find the protein-protein(p-p) interactions(Λ) and assume that we have λ_1 (cell-cycle control p-p interaction), λ_2 (signal transduction p-p interaction), λ_3 (vesicular transport p-p interaction), λ_4 (cell polarity p-p interaction), etc. Because all the λ_n, such as the interactions in the aspects of cell-cycle control, signal transduction, vesicular transport, cell polarity, RNA-processing/modification, RNA-splicing, mitosis, protein synthesis, carbohydrate metabolism, cell stress, chromatin/chromosome structure, protein folding, protein degradation, amino acid metabolism, etc, belong to Λ (p-p interaction), ie $\lambda_1, \lambda_2, \lambda_3, \lambda_4 \cdots\cdots \in \Lambda$, then all the smaller relation networks form a total and larger "protein-protein interaction network". That is $(\lambda_1, \tau, o, \gamma) \cup (\lambda_2, \tau, o, \gamma) \cup (\lambda_3, \tau, o, \gamma) \cup (\lambda_4, \tau, o, \gamma) \cup \ldots\ldots \in (\Lambda, \tau, o, \gamma)$.

Ⅴ. Very Ambitious Plans of Biomedical Integration

Ⅵ. The More Than Ten Years Long Dream

As stated above, the goal of this series of articles is to explore a new research virgin land for Biomedical Informatics, i.e. to see how widely, how precisely, to what a degree, we can integrate the existing biomedical data or knowledge. The original idea has come from not only the high integrality, ie a basic nature of organisms or the objects which Biology and Medicine deal with, but

also from the desire to promote a new scientific spirit for scientists paying more attention to more aspects of biomedical knowledge rather than limiting themselves life-long to their narrow domains only. Furthermore, BMKI is doubly parallel to the current hot area of biomedical knowledge engineering such as various biomedical ontologies and their integration in future, digital or virtual human, etc.

The author has devoted himself to the research on the simulation and integration of medical knowledge since 1989[17-20] and advanced in one of his articles the "new research on large-scale integration of biomedicine" in 1991[21]. Its activities since then have gone through several exploratory steps of theoretical discussions on Quantitatively or Qualitatively Medicine Simulating and Operating by Computer(QMSOC)[17-20], Homo-Information Coded Editing(HICE) technique[22], Integrateable Relationized Medical Electronic Book (IRMEB)[23], Human-Computer Commonly Understandable and Communicatable Medical Language（or Human-Computer Shareable Language, HCSL）[24],and the development of Internet web site for IREMB(having been thoroughly destroyed for three times by the ISP without any explanations). Because of less understanding and supports, for more than ten years, the author has acted as the theory explorer, algorithm designer, data and knowledge organizer and the operative program and AI program developer, and web site master as well. That is why the author laughed at himself as doing certain "single-gun-and-horse huge project". May be in that case, any one in the world cannot help feeling alone and hopeless.

The things, however, have dramatically been changed since the popularization of Internet, where the so many scientists publish their own best works. The Human Genome Project(HGP) which was finished in last year is doubtlessly a great work in the history of life sciences. After HGP, people has realized more clearly that it is hardly possible to understand the true essence by the separate information of genes and proteins(e.g. enzymes). In the coming time, ie so called meta-genomic time, the integration researches such as protein network integration, drug ontology synthesis, biochemistry reaction network, etc are going to be merging one after another.

Suddenly BMKI found so many friendly researches and sciences and is getting excited. With best wishes!

（to be continued）

References

1. JH van Bemmel, MA Musen 主编，包含飞，郑学侃主译：《医学信息学》，上海：上海科技出版社，2002

2.郝柏林，张淑誉：《生物信息学手册》，上海：上海科技出版社，2000

3. 张镜如主编，乔健天副主编：《生理学》，北京：人民卫生出版社，86-88，39-45，1996

4. 包含飞：续议中医学是复杂性科学——中医标准化预备研究之三（待发表）

5.Bao H.F.: HCSL:A Human-Computer Commonly Understandable and Communicatable Medical Language, 《Proceedings of The First China-Japan-Korea Joint Symposium on Medical Informatics(CJKMI'99)》, p177-181,1999

6. 包含飞:创造一种人机共解互通的医学语言(西医学部分), 第八届全国医药信息学大会论文集, CMIA′99(电子工程师增刊), 1999, 98:18-24

7．刘沁，包含飞：人机共享语言 HCSL（医学）的第二版的计算机实现，中外名医杂志，（1）：78-80，2001

8 . UMLS Knowledge Sources 14 Edition, January Release, http://www.nlm.nih.gov/research/umls/UMLSDOC_2003AA.pdf

9．Benno Schwikowski, Peter Uetz and Stanley Fields: A network of protein–protein interactions in yeast，http://itgmv1.fzk.de/www/itg/uetz/publications/Schwikowski2000.pdf

10． Dieter Fensel, Mark A. Musen : The Semantic Web: A Brain for Humankind，http://www.cs.umbc.edu/771/papers/ieeeIntelligentSystems/introduction.pdf

11．Klaus Prank：Data mining and mathematical modeling in systems biology，www.systemsbiology.org

12．Web site: Entelos is the leader in predictive biosimulation. http://www.entelos.com/science/index.html

13. Nick Monk, Philip Maini, Denis Noble, Tim Pedley and Michael Akam: Vertical Integration in Bioloy: From Molecules to Organisms, http://www.newton.cam.ac.uk/programs/ICB/icbw01.html

14.Stanford University: The Bio-X program, http://cmgm.stanford.edu/biochem/biox/intro.html

15.Cambridge Biology Integration Group(BIG): http://www.gen.cam.ac.uk/~big/

16.Biology of Ageing e-Science Integration and Simulation system (BASIS): http://umbriel.dcs.gla.ac.uk/NeSC/action/projects/project_action.cfm?title=91 or http://www.basis.ncl.ac.uk/index.html

17.Bao H.F.: The structure characteristics of the new research QMSOC and its relevant operators, J Tongji Med Univ, 1989, 9(4): 235-238

18.Bao H.F.: Quantitative and Computerized Medicine--New research QMSOC(I), The proceeding of the First Conference of the Frontier for Life Sciences on Central-south China, 1989, 211-216

19.Bao H.F.: The quantitative integration of biomedical information by computer--New research QMSOC(II), The Proceeding of the First Conference of the Frontier for Life Sciences in Central-south China, 1989, 216-222

20.Bao H.F.: The new functions of Quantitatively Medicine Simulating and Operating by Computer-New research QMSOC(III), 1990, J Tongji Med. Univ, 10(1): 52-56

21.Bao H.F., Geng J.H. and Su Z.F.: Pansystems Methodology(PM) and a new research on large-scale integration of biomedicine—An introduction of QMSOC and its recent progresses , Journal of Jiangsu Institute of Technology, 1991, 4 (2): 69-75

22. 包含飞，刘庚妹：计算机编码编辑技术（HICE）及其初级信息开发功能——新研究 QMSOC(V)，计算机应用研究（1991 年专辑第 2 号），1992，4-9

23.Bao H.F., Ni X.W., Lou S.: Integratable Relationized Medical Electronic Book (IRMEB)--- An Exploration of An New Type of Intelligent Knowledge Medium Under the Influence of Pansystems Theory. Advances in Systems Science and Applications (Inauguration Issue) 1995,p304-309

24. 包含飞：创造一种人机共解互通的医学语言(西医学部分)，第八届全国医药信息学大会论文集，CMIA′99(电子工程师增刊)，1999，98:18-24