KEGG Slide Show: Part 1

IS GENOME A BLUEPRINT OF LIFE ?

A Philosophical Background of KEGG

Minoru Kanehisa
Institute for Chemical Research
Kyoto University

Excerpts from the talks given at:

NATURE's first International Conference in Korea: Bioscience Meets Engineering
November 3, 1997 (Seoul, Korea)

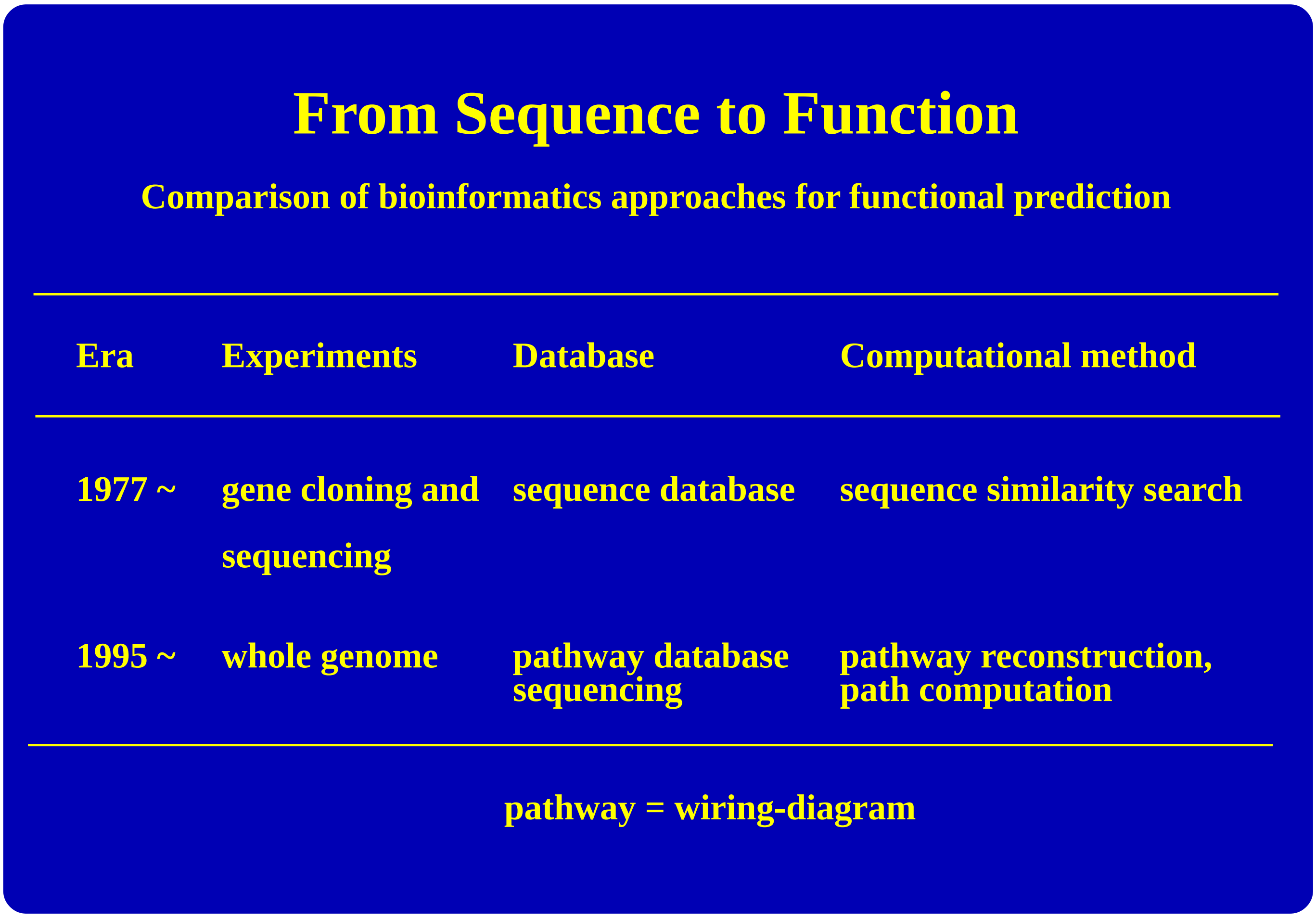
Yokohama Forum for the 21st Century: New Frontiers of Biosciences Created by Genome Biology
November 4, 1997 (Yokohama, Japan)

Japan-Germany Forum on Information Technology
November 11, 1997 (Nagano, Japan)

Nihon Silicon-Graphics Cray Bioinformatics Seminar
November 19, 1997 (Tokyo, Japan)

Philippines National Academy of Science and Technology: The Third D.L. Umali Memorial Lecture
December 3, 1997 (Los Baños, Philippines)

FAOBMB 25th Anniversary Symposium
December 4, 1997 (Manila, Philippines)



From Sequence to Function

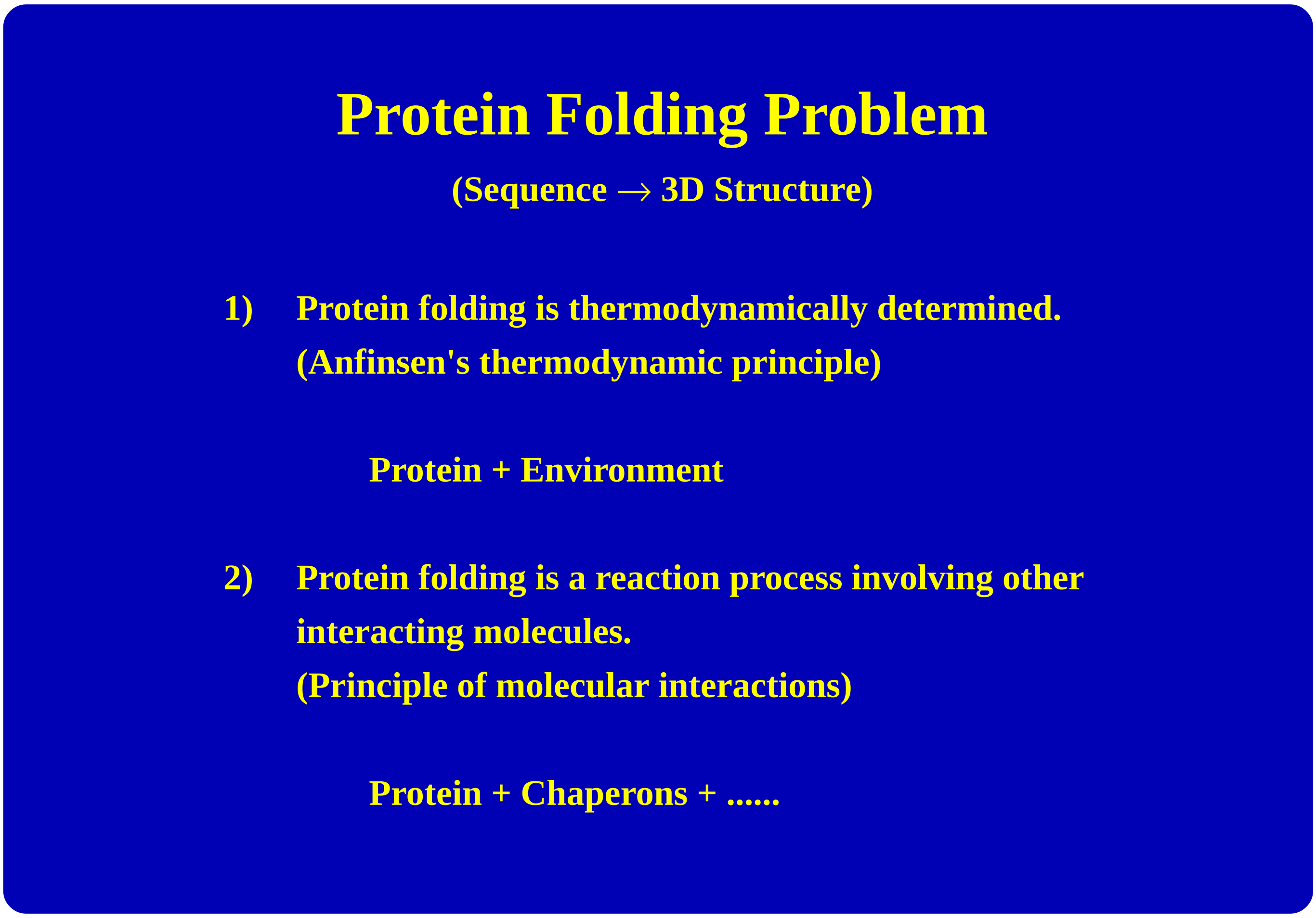
Comparison of bioinformatics approaches for functional prediction

Era	Experiments	Database	Computational method
1977 ~	gene cloning and sequencing	sequence database	sequence similarity search
1995 ~	whole genome	pathway database sequencing	pathway reconstruction, path computation

pathway = wiring-diagram

In 1977 a tiny virus genome *ϕx174* was sequenced by Frederick Sanger and his group, which was the beginning of the era of gene cloning and sequencing where searching for a specific gene and determining its sequence have become a most fundamental approach in all areas of biological sciences. In 1995 the first complete genome of a free living organism, *Hamophilus influenzae*, was sequenced by Craig Venter and his group, which is considered the beginning of the era of whole genome sequencing. Sequence information was the starting point of understanding the function of a single gene or a single gene product before, but now it is the starting point of understanding the entire mechanisms of biological behaviors in living organisms.

One of the major tasks of bioinformatics is to develop new databases and new computational technologies that will help understand the biological meaning encoded in the sequence data. The collection of all known sequences in the form of a sequence database and the similarity search against it have been extremely useful in predicting biological functions of individual genes and individual proteins. Then, in the era of whole genome sequencing what kind of databases and computational technologies do we need to develop for predicting systems behaviors of a network of genes and a network of proteins?



Protein Folding Problem

(Sequence → 3D Structure)

1) **Protein folding is thermodynamically determined.**
   (Anfinsen's thermodynamic principle)

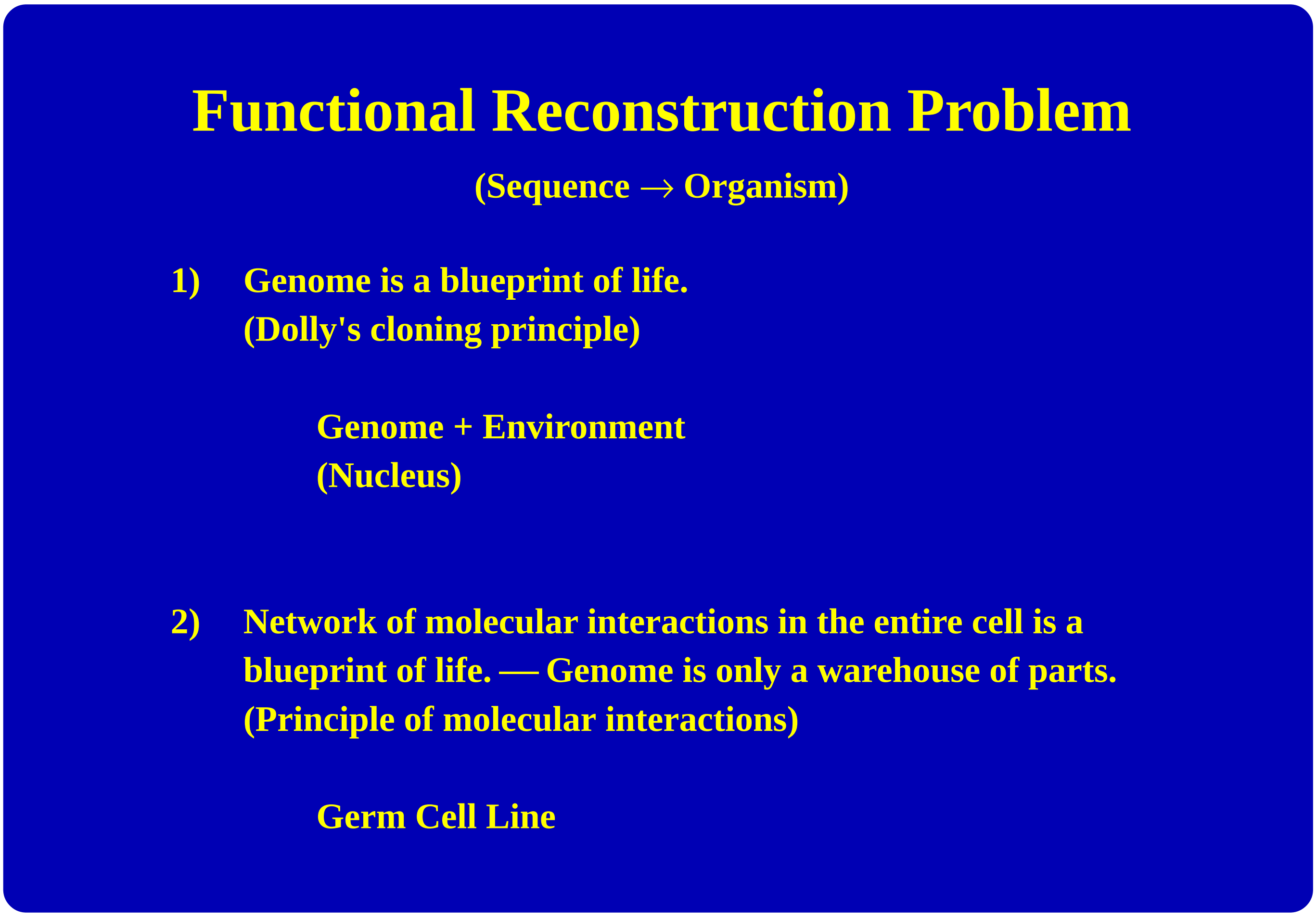
   Protein + Environment

2) **Protein folding is a reaction process involving other interacting molecules.**
   (Principle of molecular interactions)

   Protein + Chaperons + ...

The so-called protein folding problem has been one of the grand challenges in molecular biology for more than 30 years. The problem is the following. Given an amino acid sequence, can you predict computationally the native three-dimensional structure of a protein? First, most people thought that the amino acid sequence contained all information that was necessary to fold up the correct three-dimensional structure, because protein folding was apparently thermodynamically determined; namely, given a proper environment a protein would fold up spontaneously. This is called Anfinsen's thermodynamic principle.

However, according to more recent experiments, protein folding is a more complex and dynamic process involving a number of other molecules, such as chaperons. The environment has to be treated as consisting of specific interactions with specific molecules rather than a smooth thermodynamic environment. This I would call the principle of molecular interactions.



Functional Reconstruction Problem

(Sequence → Organism)

1) **Genome is a blueprint of life.**
(Dolly's cloning principle)

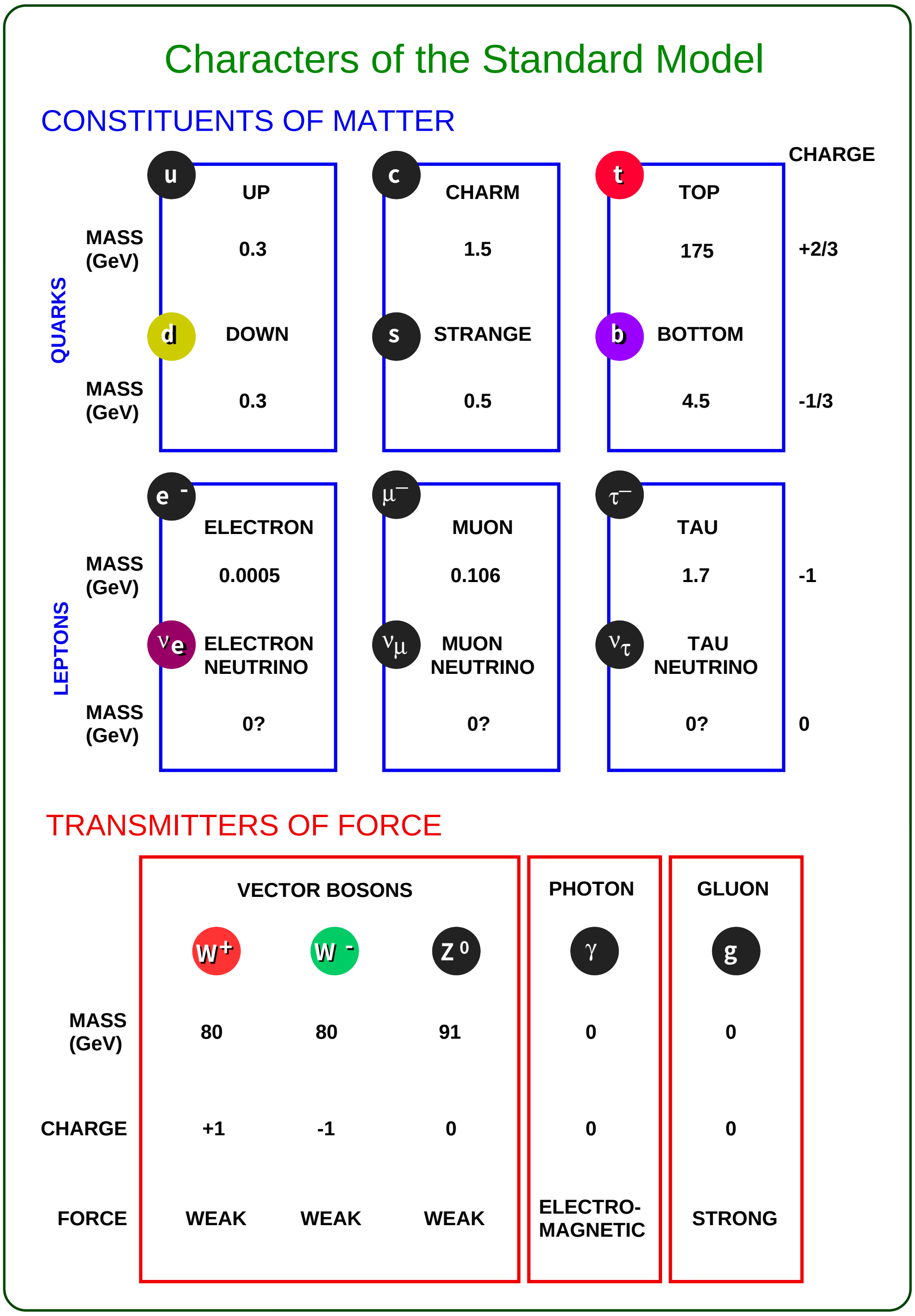
Genome + Environment
(Nucleus)

2) **Network of molecular interactions in the entire cell is a blueprint of life. — Genome is only a warehouse of parts.**
(Principle of molecular interactions)

Germ Cell Line

In the era of whole genome sequencing, we are faced with another grand challenge problem, which may be called the functional reconstruction problem. The problem is the following. Given an entire genome sequence, can you predict computationally the systems behavior of a biological organism? Here again, a traditional view is that the genome is a blueprint of life containing all necessary information that would make up a biological organism. If you replace a nucleus, then you get a clone. So, this might be called Dolly's cloning principle.

However, I think an alternative view can also be taken where the genome is a warehouse of parts, or building blocks of life, and all the regulatory signals in the genome are simply bar codes to retrieve them. In this view a blueprint of life is written in the entire cell as a network of molecular interactions. Unless you start with an exactly identical germ cell containing an exactly identical nucleus, you cannot make a clone. After all, we inherit from our mother not just a nucleus, but an entire cell. Thus, this is again the network of molecular interactions.



Characters of the Standard Model

CONSTITUENTS OF MATTER

QUARKS					CHARGE
		u UP	c CHARM	t TOP	
	MASS (GeV)	0.3	1.5	175	+2/3
		d DOWN	s STRANGE	b BOTTOM	
	MASS (GeV)	0.3	0.5	4.5	-1/3

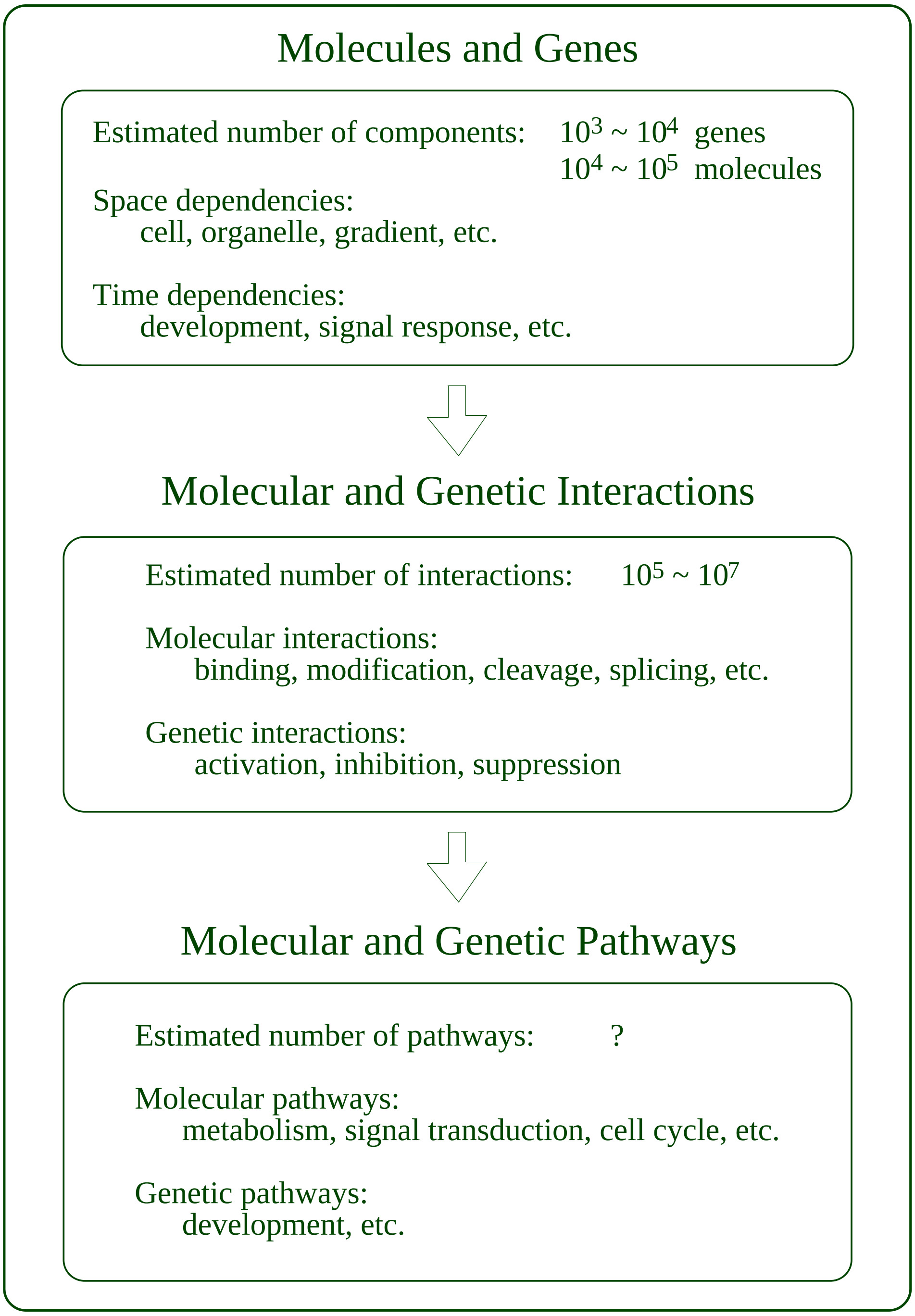
LEPTONS					
		e^- ELECTRON	μ^- MUON	τ^- TAU	
	MASS (GeV)	0.0005	0.106	1.7	-1
		ν_e ELECTRON NEUTRINO	ν_μ MUON NEUTRINO	ν_τ TAU NEUTRINO	
	MASS (GeV)	0?	0?	0?	0

TRANSMITTERS OF FORCE

	VECTOR BOSONS			PHOTON	GLUON
	W^+	W^-	Z^0	γ	g
MASS (GeV)	80	80	91	0	0
CHARGE	+1	-1	0	0	0
FORCE	WEAK	WEAK	WEAK	ELECTRO-MAGNETIC	STRONG

Liss, T.M. and Tipton, P.L.; The Discovery of the Top Quark.
Scientific American, September 1997

If you are familiar with elementary particle physics, you know that there are two classes of elementary particles. Constituents of matter, or building blocks of nature, form one class of elementary particles and transmitters of force, or interactions between building blocks, form another class of elementary particles. Therefore, both building blocks and interactions are the most fundamental aspects of nature. I believe that building blocks and interactions are also the most fundamental aspects of life.



Molecules and Genes

Estimated number of components: $10^3 \sim 10^4$ genes
$10^4 \sim 10^5$ molecules

Space dependencies:
cell, organelle, gradient, etc.

Time dependencies:
development, signal response, etc.

Molecular and Genetic Interactions

Estimated number of interactions: $10^5 \sim 10^7$

Molecular interactions:
binding, modification, cleavage, splicing, etc.

Genetic interactions:
activation, inhibition, suppression

Molecular and Genetic Pathways

Estimated number of pathways: ?

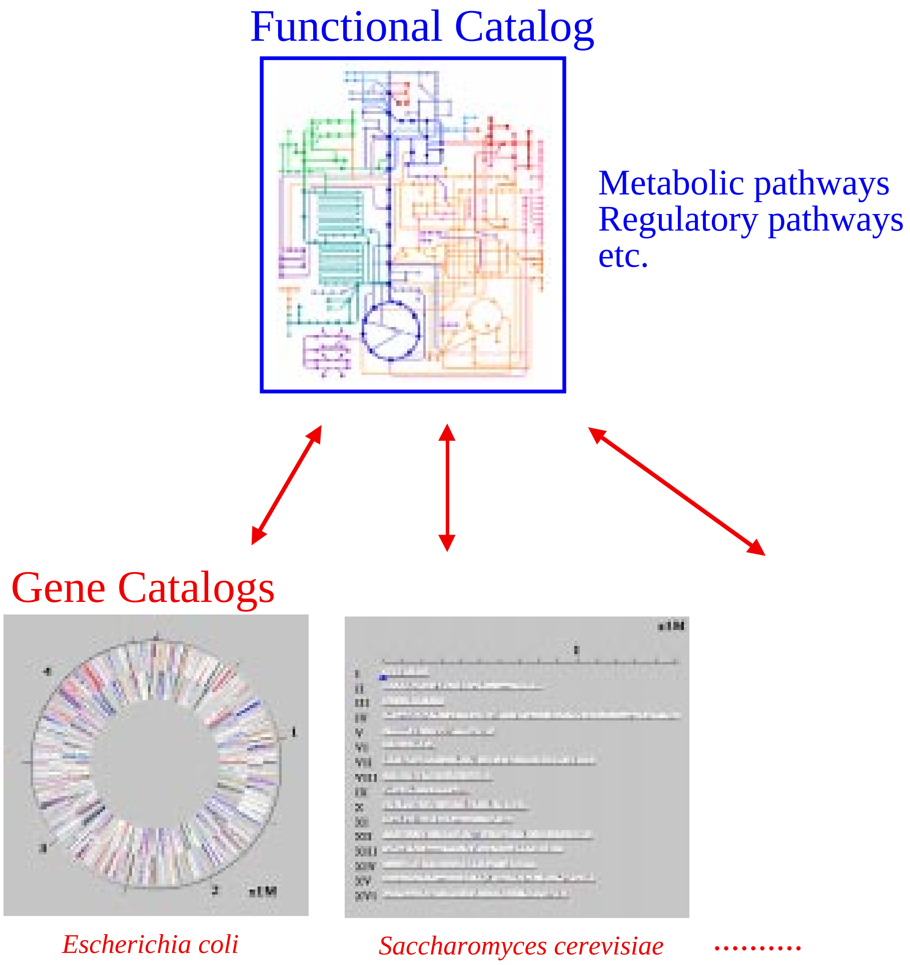
Molecular pathways:
metabolism, signal transduction, cell cycle, etc.

Genetic pathways:
development, etc.

The building blocks include not only the biological macromolecules, proteins and nucleic acids, but also other molecules such as small chemical compounds and metal ions. Whether the genome is a blueprint or not, it is impossible from a practical point of view to fully make sense out of the sequence data without additional information. For example, the function of at least one third of the 6,000 genes identified in the yeast genome is still unknown. In order to obtain any functional clue, systematic gene disruption experiments are under way to see the effect of knocking out a gene in the rest of the genes, or to detect a gene-gene interaction. Also there are systematic protein-protein interaction experiments where the binding patterns of 6000 by 6000 yeast proteins are examined by the yeast two-hybrid system.

We believe that traditional experiments in genetics, biochemistry, and molecular and cellular biology have contributed to the accumulation of a huge body of knowledge on molecular interactions and resulting molecular pathways in selected aspects of living systems. The problem is that it is not computerized.

From Gene Catalog to Functional Catalog
The KEGG Project



Functional Catalog

Metabolic pathways
Regulatory pathways
etc.

Gene Catalogs

Escherichia coli

Saccharomyces cerevisiae

....

In 1995 we initiated the KEGG Project under the Japanese Human Genome Program in order to fully utilize the biological knowledge on interactions and pathways, as well as the interaction data generated by new systematic experiments, KEGG stands for Kyoto Encyclopedia of Genes and Genomes.

The primary objective of KEGG is to computerize the current knowledge of molecular pathways; namely, metabolic pathways and a number of regulatory pathways including signal transduction, cell cycle, and developmental pathways. At the same time, KEGG maintains gene catalogs for all the organisms that have been sequenced and links each gene product to a component on the pathway. Because we need an additional catalog of building blocks, KEGG also organizes a database of all chemical compounds in living cells and links each compound to a pathway component. And finally, KEGG aims at developing new bioinformatics technologies toward functional reconstruction.