Global Focus on Knowledge Lecture Series
Rapidly Developing Life Sciences:
Puzzling out the Systems of Life
13 May 2009

Part 4 in the Genomes, Information and Evolution Series
From Genomes to an Understanding of Life Systems

Database Center for Life Science, Research Organization of Information and Systems
Department of Computational Biology, Graduate School of Frontier Science, The University of Tokyo
Center for Information Biology and DNA Databank of Japan, National Institute of Genetics

Toshihisa Takagi

Review Through Previous Lecture (Part 3)

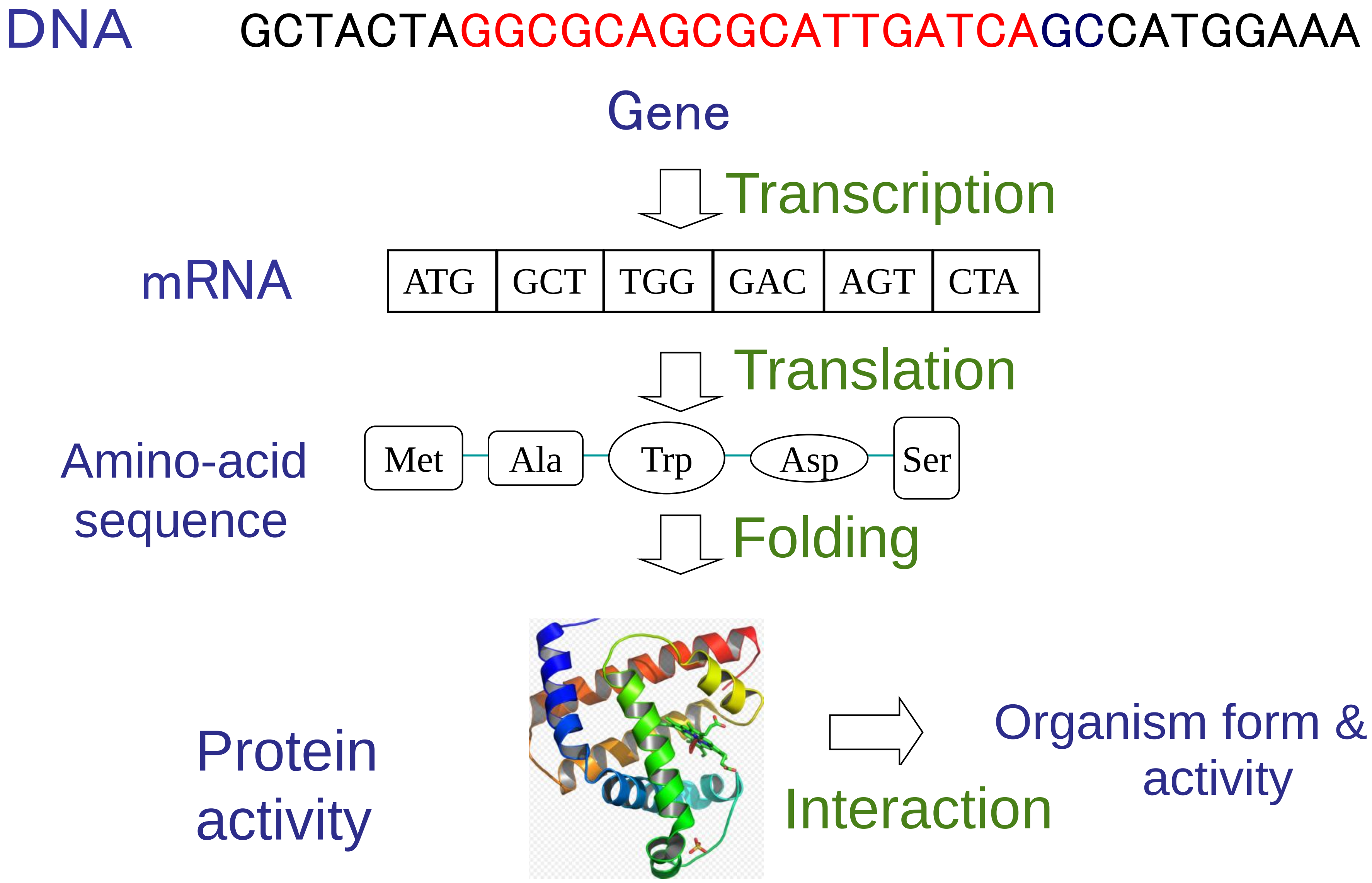
- We have learned a great deal from sequencing genomes (i.e. finding their ATCG sequences) and analyzing them with fast computers and algorithms:
  - What sort of genes are coded?
  - What genes are common to all life forms? What genes are unique?
  - How were these genes acquired in the process of evolution?
  - Have our theories of evolution proved adequate?
  - What about human diversity? How has it evolved?

Can We Say That We Understand Life Now?

- The genome is the blueprint of life, its recipe.
- So knowledge of genomes is understanding of life?
- Can we prevent, diagnose and treat diseases, and develop drugs for them?
- Can we grow rice that tastes good and that withstands cold and pests?

- The answer is No.
- Why? Because we lack the knowledge to interpret the blueprint, or recipe, of life.
- And because genes are elaborately entwined in the way they manifest their functions. That is, they comprise a system.

Genome as Blueprint of Life



DNA GCTACTAGGCGCAGCGCATTGATCAGCCATGGAAA

Gene

Transcription

mRNA

ATG GCT TGG GAC AGT CTA

Translation

Amino-acid sequence

Met Ala Trp Asp Ser

Folding

Protein activity

Interaction

Organism form & activity

Today's Lecture

- Previous lectures covered what we have learned and what seems knowable
- Today is about what we still don't know, what needs study and research
- Researchers laboring at the cutting edge of research
- Post-genome (post-sequencing) trends:
    - Understanding systems
    - Omes, omics and systems biology
    - Databases and integration
    - Ontologies and text mining
- The spread of genome research and research based on genome foundations
- A paradigm shift in life science research
- The need for computational biologists: What we want from you

What Is It to Understand Life Systems?

- What do we mean by "system"?
  - Does not translate well to Japanese
  - A useful word
  - One seems to understand the word, but does not
  - Definition varies subtly from field to field, person to person

- *Our* definition of "system" will be a set combining varied components that as a whole performs complex operations (ones that could not be inferred from a catalog of its components).
- Life is a system comprising such components as DNA (genes), RNA, proteins and chemical compounds.

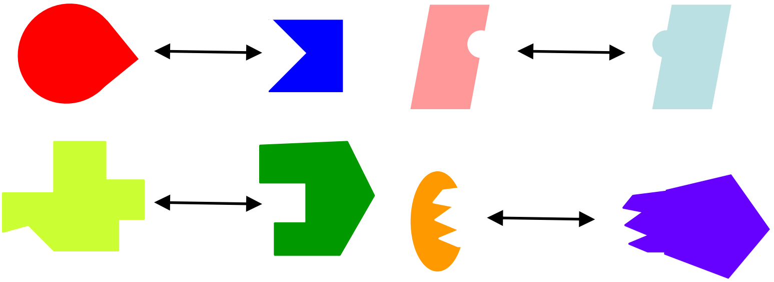
Different Levels of Understanding

- Listing and cataloging components
- Describing the forms of individual components
- Identifying where and when components are used
- Describing the operation of individual components
- Describing component relationships and networks
- Describing the operations and properties of sub-networks
- Predictability of system behavior
- Controllability of system behavior
- Ability to experiment with identical systems

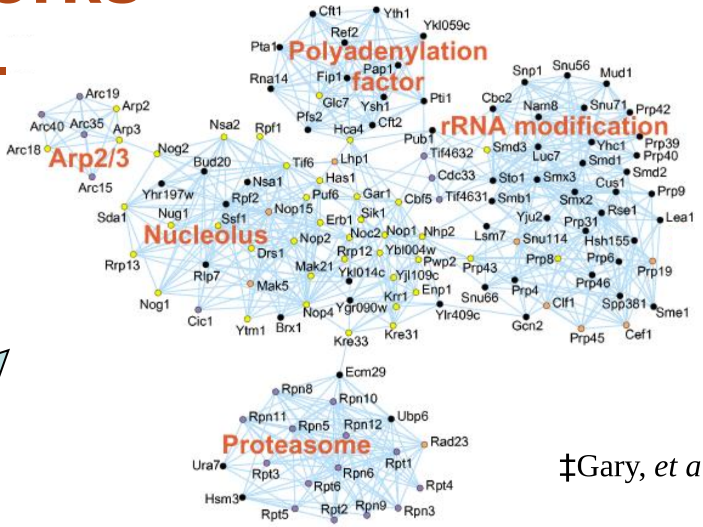
Omes, Omics and Systems Biology

- Identifying, listing and cataloging components -- Genomes, genomics
- Describing the forms of individual components -- Proteomes, proteomics
- Identifying where components are used -- Transcriptomes
- Describing the operation of individual components -- Functional genomics
- Describing component relationships and networks -- Interactomes
- Describing the operations and properties of sub-networks -- Network biology
- Predictability of system behavior -- Systems biology
- Controllability of system behavior -- Systems biology
- Ability to experiment with identical systems -- Synthetic biology

Analysis of Interactive Networks

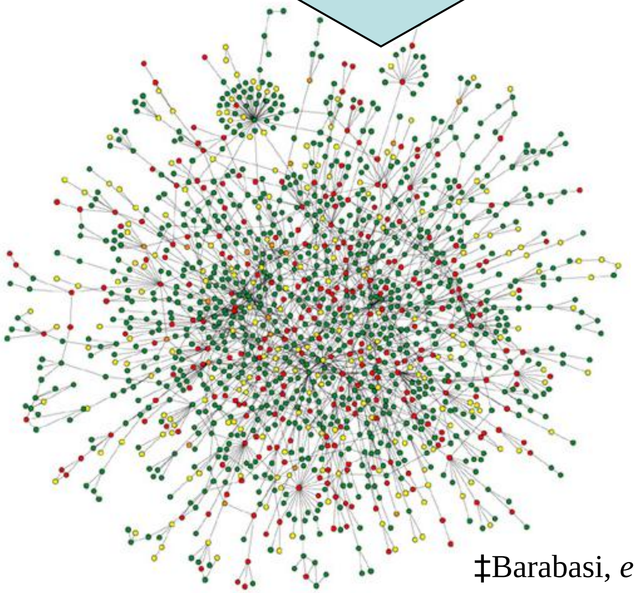


Join up individual interactions



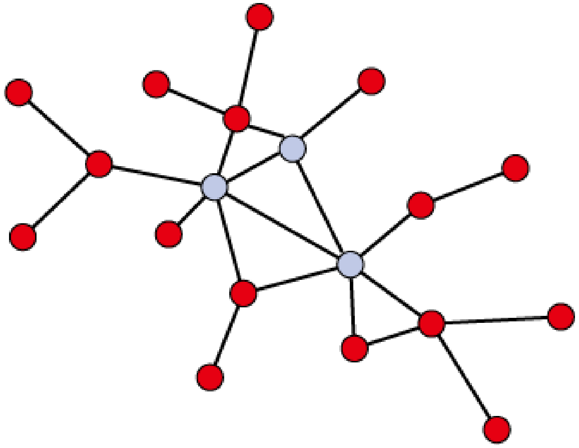
‡Gary, *et al*. 2002

Identify functional modules



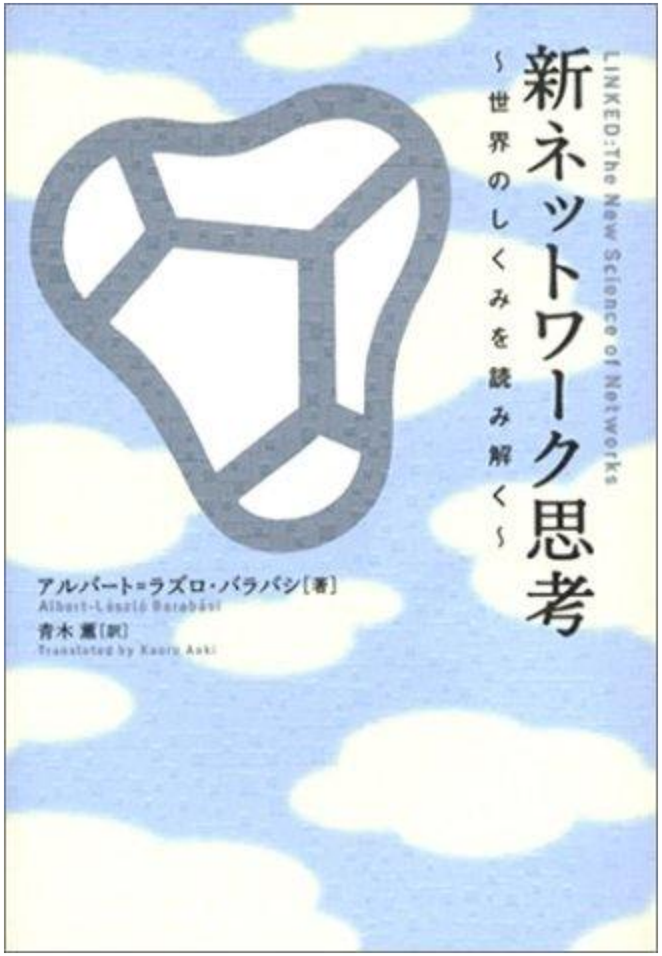
‡Barabasi, *et al*. 2001

Interactive networks properties



Analyze network

Describing the Common Properties of Networks



‡NHK(2002)

- Albert-László Barabási, physicist
- Common properties of networks
  - Protein networks
  - Metabolic networks
  - Computer networks
  - Social networks
  - Small-world networks
  - Scale-free networks

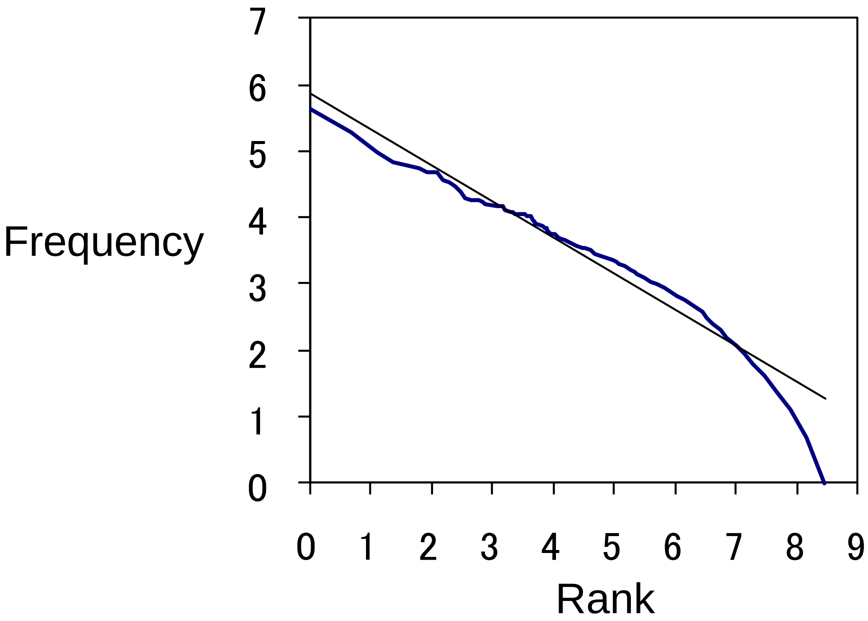
Examples of Network Biology

- Zipf's law
- Frequency of English words

•the	10%
•of	5%
•and	3%
•to	2%

Gene	Interactions
JSN1	282
SRP1	197
NUP116	147
⋮	⋮
PBN1	1

Yeast interactions: DIP, 2003



What We Would Like to Know

- Relationship between structure and function
- Relationship between genotype and phenotype

- Move from an understanding of one-to-one relationships to an understanding of *n*-to-*n* relationships
- What is the relationship (system) of *n* elements?

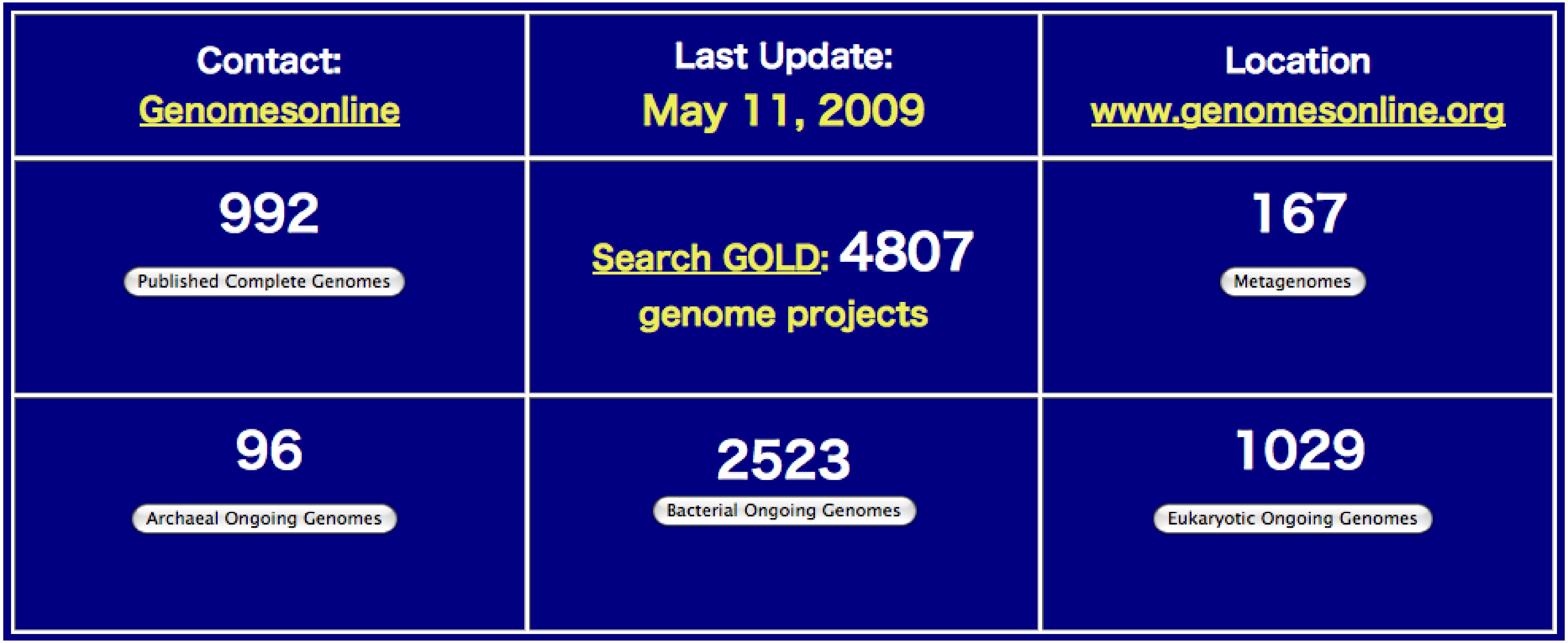
- We would like to know these things at each of level of the hierarchy of life (cells, organs, individuals, populations).

Significance of Genomes in Life-Sciences Research

- A genome is a list of components.
- Finiteness and completeness allow us to refine and reduce the search space.
- Genomic information and analytical techniques serve as the foundations of many different exhaustive measurements (e.g. DNA chips).

- Genomes permit the integrated study, in conjunction, of universality and diversity.
- Note that the genome is different from the other omes.
    - Post-post-genome (sequencing) we return again to the genome (sequencing)

Where Genome Analysis Stands Today



Contact: Genomesonline	Last Update: May 11, 2009	Location www.genomesonline.org
992 Published Complete Genomes	Search GOLD: 4807 genome projects	167 Metagenomes
96 Archaeal Ongoing Genomes	2523 Bacterial Ongoing Genomes	1029 Eukaryotic Ongoing Genomes

Source: GOLD (Genome Online Database) v.2.0

Genome Size and Gene Count

	Size(bp)	Genes
Mycoplasma	0.6mn	500
E. coli	4.6mn	4,400
Budding yeast	12mn	6,100
Fission yeast	13mn	4,900
Slime mold	34mn	16,000
Nematode	100mn	20,000
Fly	130mn	14,000
Blowfish	340mn	28,000
Chicken	1100mn	25,000
Mouse	2600mn	19,000
Human	2900mn	22,000

Disjunction of Gene Count and System Complexity

- The great surprise and mystery presented by genome research

- Research underway to explain this phenomenon includes:
    - Production of various proteins by substituting nucleotides in gene regions
    - Mechanisms of transcriptional control
    - Complex molecular interactions
    - Non-protein-coding RNA (functional RNA)
    - Epigenomes

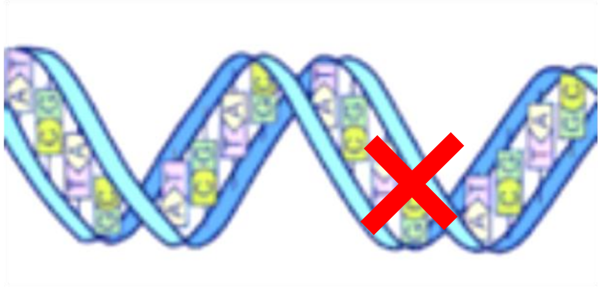
Approaches to the Understanding of Systems

- Structural data collection
  - Comprehensive collection of information on components and their relationships
  - Development of high-throughout, high-precision instrumentation

- Functional data collection
  - Use RNA interference, or other means, to inhibit gene functions and observe phenotypes
  - Survey extant literature for lapsed knowledge

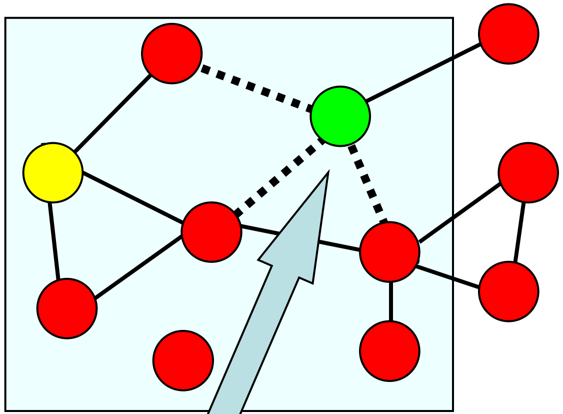
- Combine structural data and functional data
  - Consolidate all information in databases
  - Network databases

Towards Systems-Level Understanding



Map to interactions

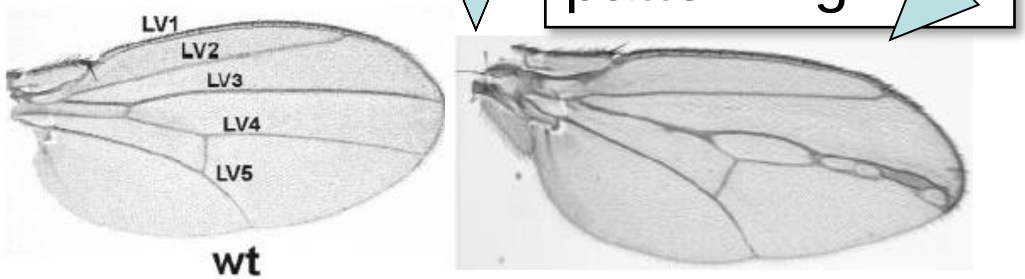
Identify functional modules



Use computers to discover hidden relationships

Gene disruption experiments

Understanding of function in wing vein patterning



‡Ascano et al., 2004

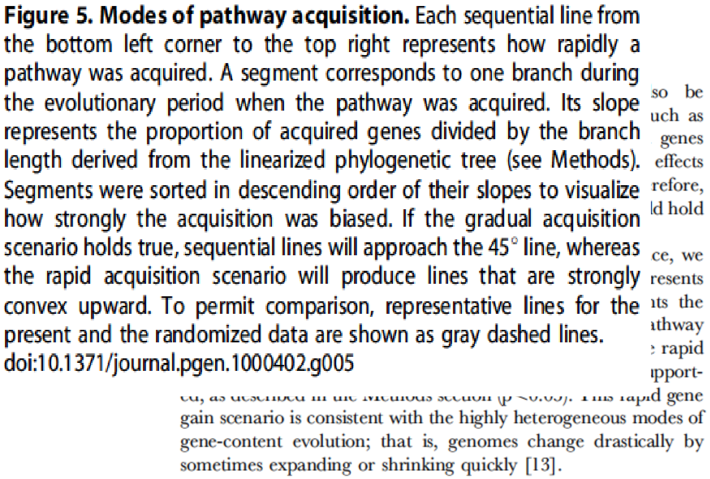


Figure 5. Modes of pathway acquisition. Each sequential line from the bottom left corner to the top right represents how rapidly a pathway was acquired. A segment corresponds to one branch during the evolutionary period when the pathway was acquired. Its slope represents the proportion of acquired genes divided by the branch length derived from the linearized phylogenetic tree (see Methods). Segments were sorted in descending order of their slopes to visualize how strongly the acquisition was biased. If the gradual acquisition scenario holds true, sequential lines will approach the 45° line, whereas the rapid acquisition scenario will produce lines that are strongly convex upward. To permit comparison, representative lines for the present and the randomized data are shown as gray dashed lines.
doi:10.1371/journal.pgen.1000402.g005

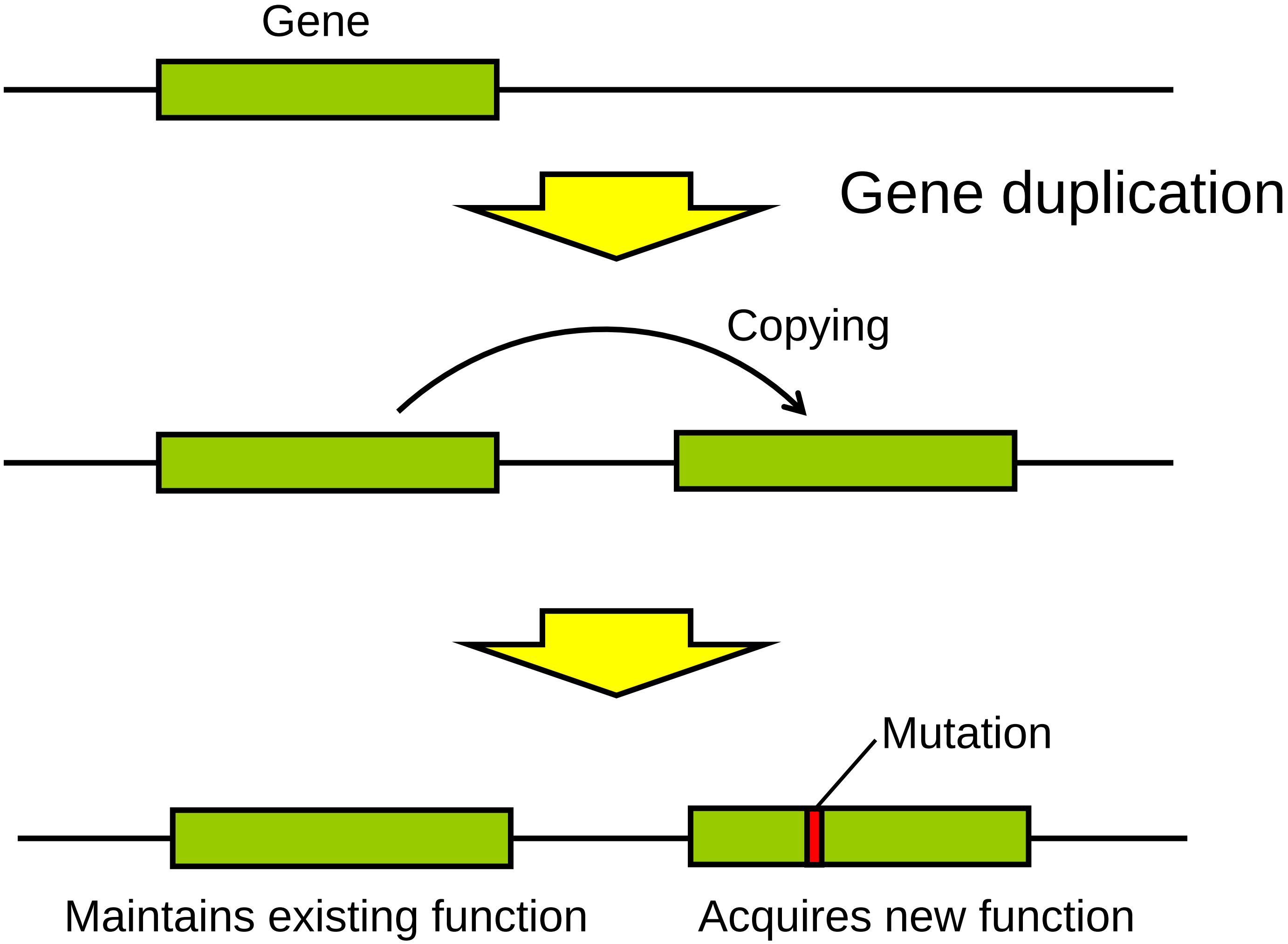
so be
uch as
genes
effects
refore,
ld hold
ce, we
resents
its the
athway
rapid
pport-
ed, as described in the Methods section (p <0.05). This rapid gene gain scenario is consistent with the highly heterogeneous modes of gene-content evolution; that is, genomes change drastically by sometimes expanding or shrinking quickly [13].

Evolutionary Analysis of Life Systems

- "Nothing in biology makes sense except in the light of evolution" -- Dobzhansky (1973)

- From evolutionary analysis of genomes to evolutionary analysis of systems
  - How did complex life systems develop?
  - Are there underlying laws that govern such development?
  - Conversely, is it possible to discover, e.g., functional modules corresponding to phenotypes particular to individual species?

Genomes Are Products of Evolution: Genomic Change is the Engine of Evolution

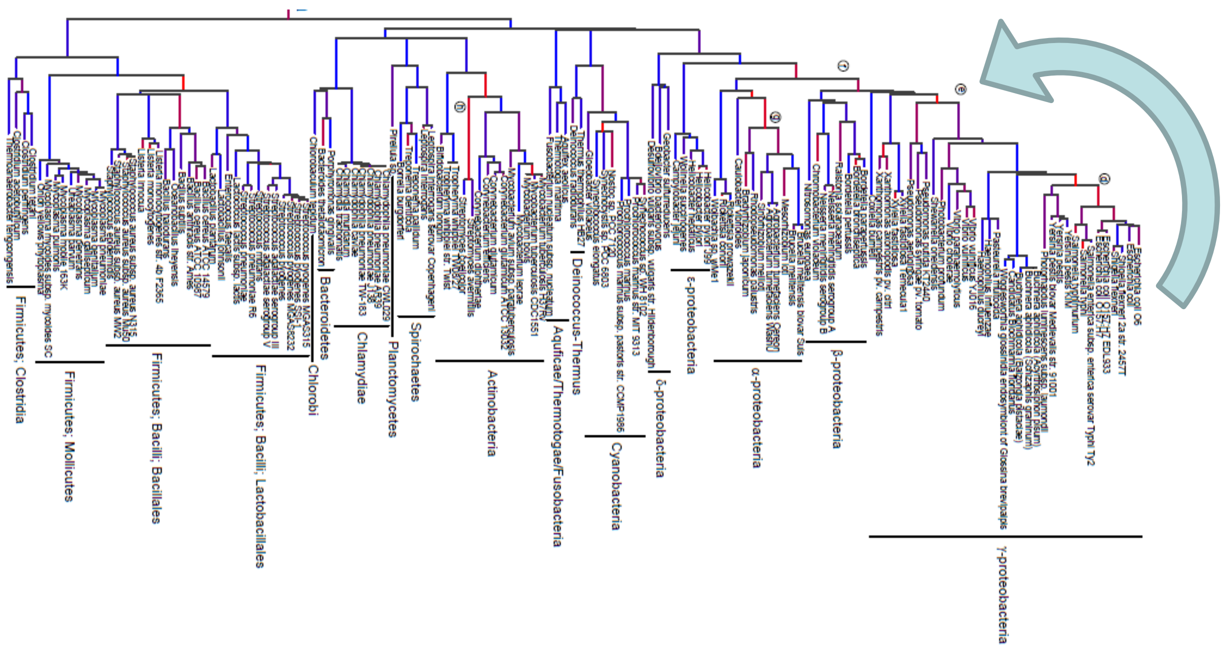


Evolutionary Analysis of Genomes

- Evolutionary analysis now possible at the genome level and with numerous species.
  - This year the number of genomes sequenced passed the 1,000 mark.

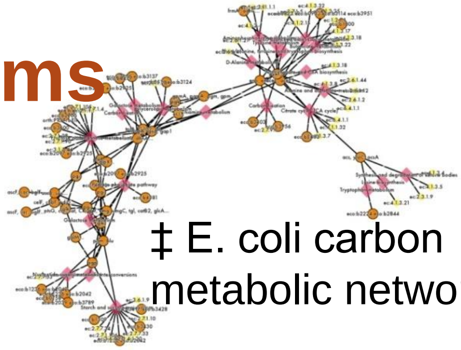
    Genome sequencing information will continue to grow explosively.

    Information on gene-expression control networks also.



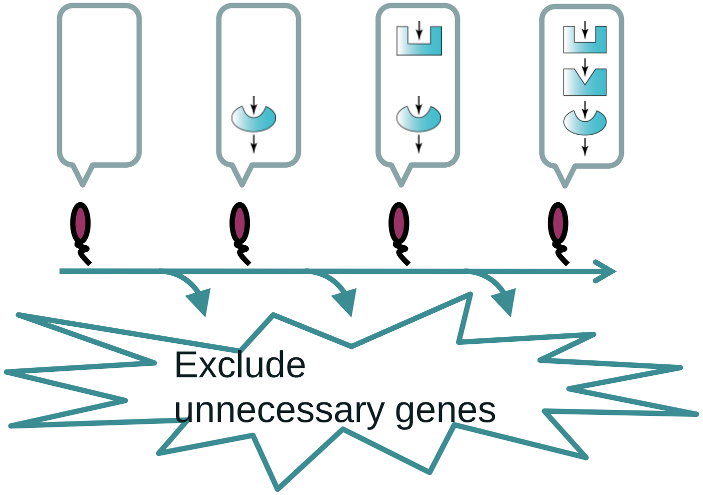
Genomic reconstruction of the evolutionary process

Evolutionary Analysis of Life Systems

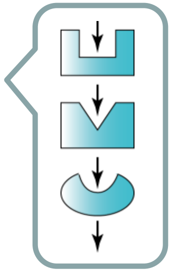


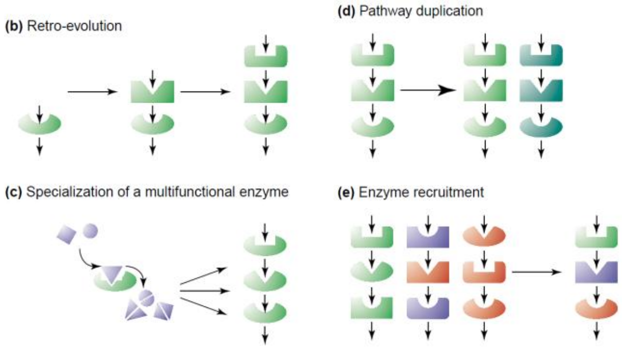
‡ E. coli carbon metabolic network

- Networks that operate through the complex coordination of numerous genes and/or proteins become useful only once all their constituent elements are in place.
  - Ex. Sequential enzymatic reactions in a metabolic network



- What is their developmental mechanism?
  - It's hard to get there by acquiring one gene at a time!
  - Construct a logical model and move the debate forward.





(b) Retro-evolution

(d) Pathway duplication

(c) Specialization of a multifunctional enzyme

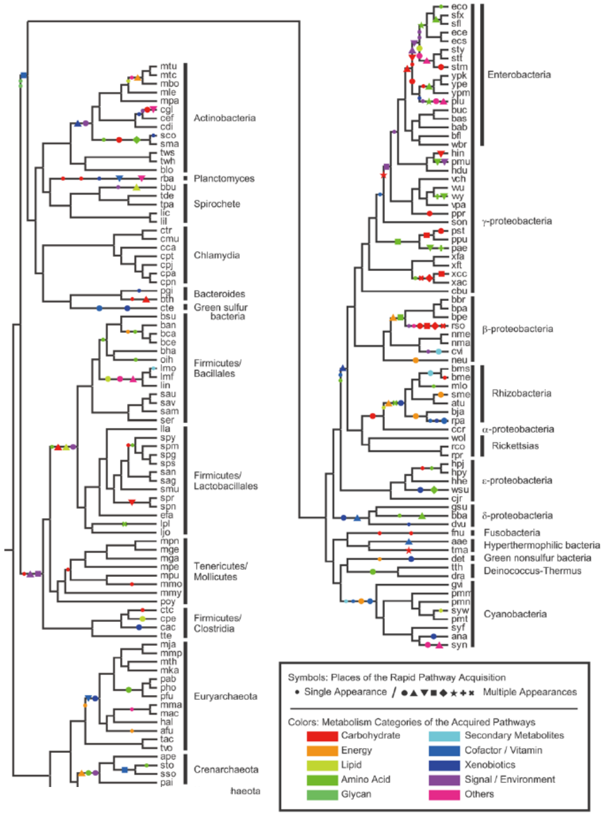
(e) Enzyme recruitment

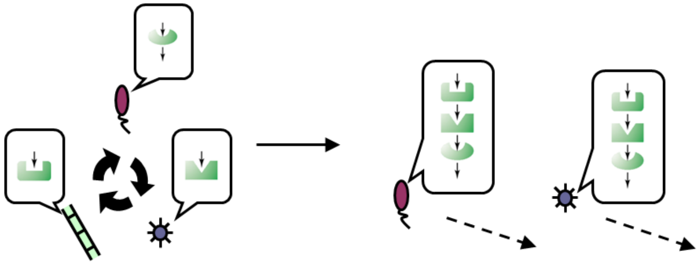
- Combine this with the evolutionary process of the genome to actually explain the developmental mechanism.

Evolutionary Analysis of Genomes and Life Systems

- Genomes of 160 prokaryotic species used to infer the genomes of extinct species
  - Metabolism databases combined to regenerate the evolutionary process of metabolic networks at the genomic and phylogenetic levels.

- The "invention" of a new module is an example of phylogenetic accomplishment at the same time and in a similar environment.
  - The evolution of networks is enabled by the ″cooperation″ of different prokaryotes in the same environment sharing genes among each other.





The Need for Databases

- Vast amounts of wide-ranging data and knowledge need to be consolidated
- High-speed computers and algorithms
- Integrated databases
- Building databases is not a chore
- An advanced intellectual activity
- Could the very act be life-science research?

- Towards database biology

How to Use Computers to Represent Complex Knowledge?

- Represent molecular-level entities
  - Genome sequences, genes --> ATGC sequences
  - Proteins --> Amino acid sequences, atomic coordinate plotting

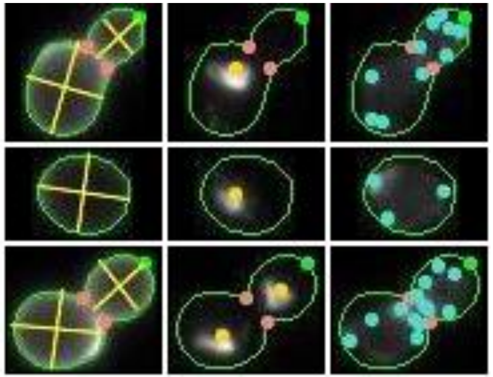
- Represent intermolecular relationships and behavior
  - Molecular interactions --> Binary relationships
  - Gene-expression information --> Intensity of expression (numeric listing)
  - Context of above data acquisition --> Ontologies

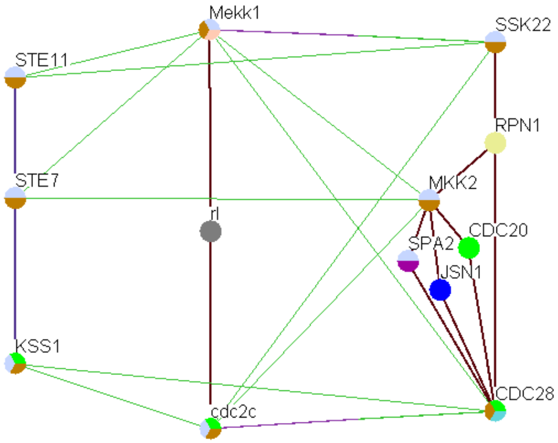
- Represent functions and phenotypes
  - Pathways, networks --> Graphs
  - Phenotypes --> Images, video, text
  - Concepts, functions and their hierarchy --> Ontologies
  - Dynamic behavior --> Mathematical models, simulators

Importance of Extensive Comparisons in the Life Sciences

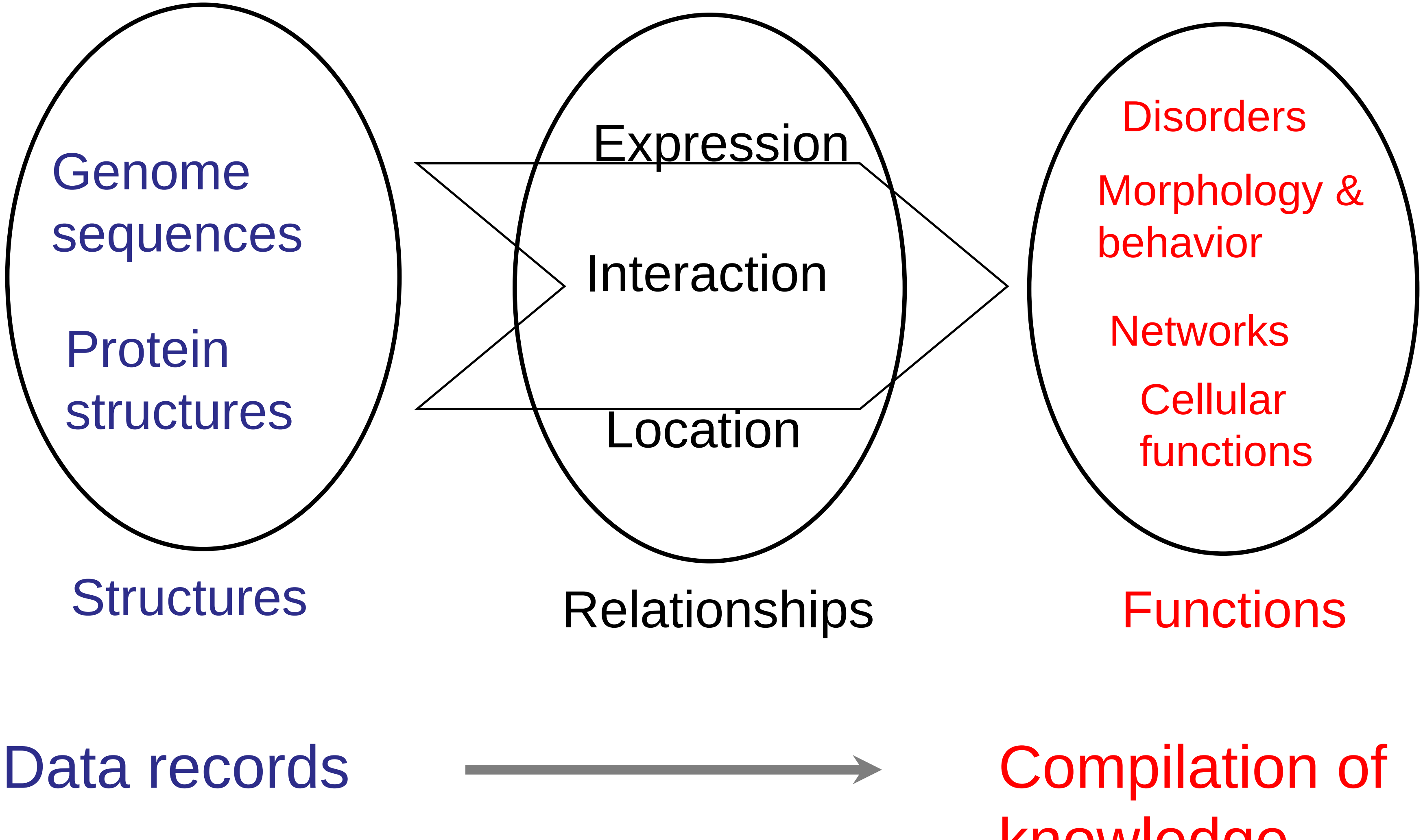
```
ATGCCTTGG-TATAATCCTATGA--GTATCG...
:  ::: :: :: :::  : :::  :: :::
AACCCT-GGGTA-AAT—TTTGAAAGTTTCGG...
```

	Liver	Heart	Brain	Lungs	Kidneys	Stomach	Intestines
Gene A	300	30	0	5	20	10	300
Gene B	0	0	10	10	200	20	30
Gene C	0	0	500	0	0	0	0
Gene D	400	20	0	10	20	10	300
Gene E	0	300	0	20	0	0	0





Trends in Database Construction



Genome sequences

Protein structures

Expression

Interaction

Location

Disorders

Morphology & behavior

Networks

Cellular functions

Structures

Relationships

Functions

Data records → Compilation of knowledge

Ontologies

- In philosophy, "ontology" refers to the study of existence.
- In computer science, it refers to knowledge-processing studies.
- In the life sciences, ontologies may comprise:
    - Gene and protein glossaries (including synonyms and abbreviations)
    - Function term standardization (across species)
    - Hierarchical relationships between different concepts and terms
    - Ways of specifying the relationships between concepts
    - Standardized database formats
- Ontologies are essential to functional databases, text mining and database integration.
- Genomes make it possible to study commonality and diversity across species.

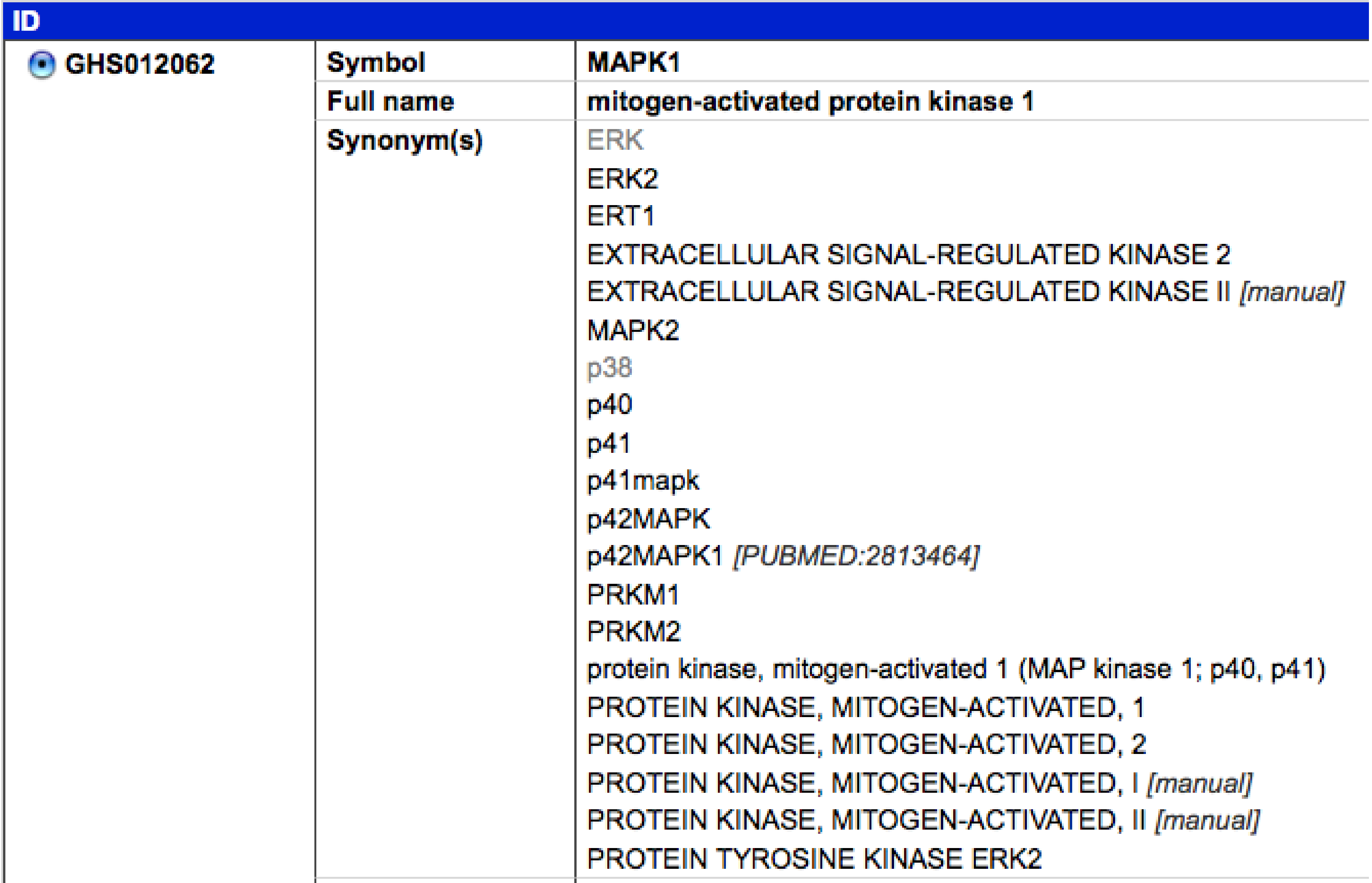
Ontologies in Genome Studies

- • In the field of biology, research communities have grown up around individual species and life phenomena.
- It is common for different communities to use different terms for the same molecules, concepts and functions (and the reverse is likewise common).

Ex. Gene definitions vary among databases.

- As genome studies require comparison and standardization across species and life phenomena, since around 2000 we have been codifying the correspondences between these terms and begun to use a standardized vocabulary.

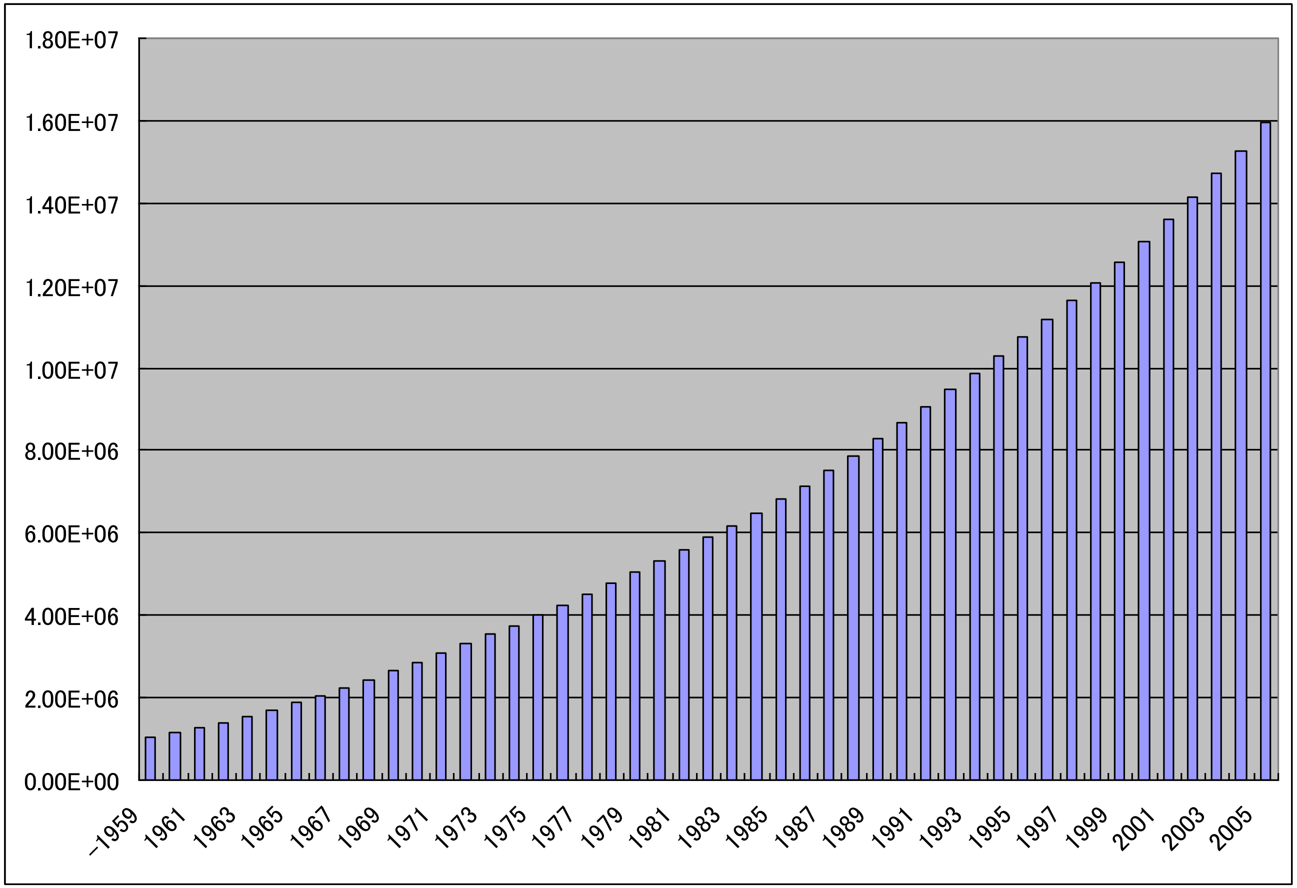
An Example of Variant Gene Naming



ID		
● GHS012062	**Symbol**	**MAPK1**
	Full name	**mitogen-activated protein kinase 1**
	Synonym(s)	ERK ERK2 ERT1 EXTRACELLULAR SIGNAL-REGULATED KINASE 2 EXTRACELLULAR SIGNAL-REGULATED KINASE II *[manual]* MAPK2 p38 p40 p41 p41mapk p42MAPK p42MAPK1 *[PUBMED:2813464]* PRKM1 PRKM2 protein kinase, mitogen-activated 1 (MAP kinase 1; p40, p41) PROTEIN KINASE, MITOGEN-ACTIVATED, 1 PROTEIN KINASE, MITOGEN-ACTIVATED, 2 PROTEIN KINASE, MITOGEN-ACTIVATED, I *[manual]* PROTEIN KINASE, MITOGEN-ACTIVATED, II *[manual]* PROTEIN TYROSINE KINASE ERK2

Growth in Publication

Abstracts found in the MEDLINE literature database



1960

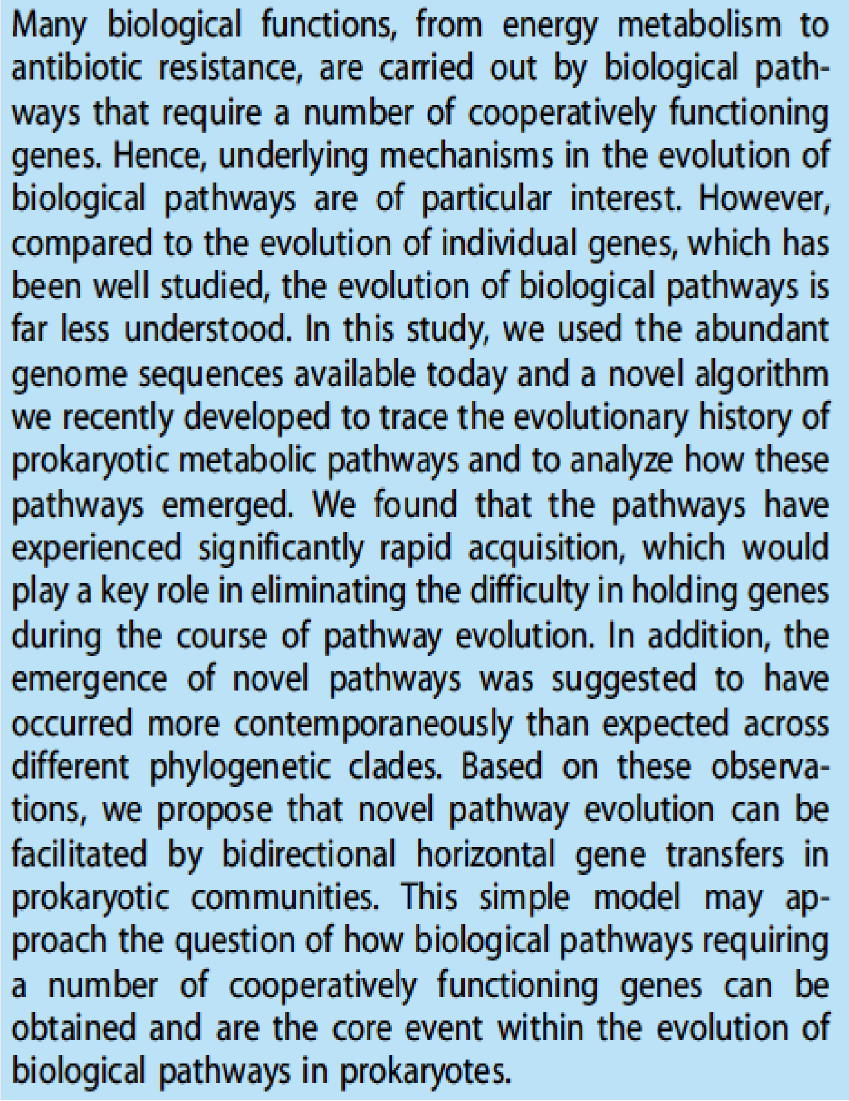
1.05million

2008

18million

Trend towards open access

Text Mining



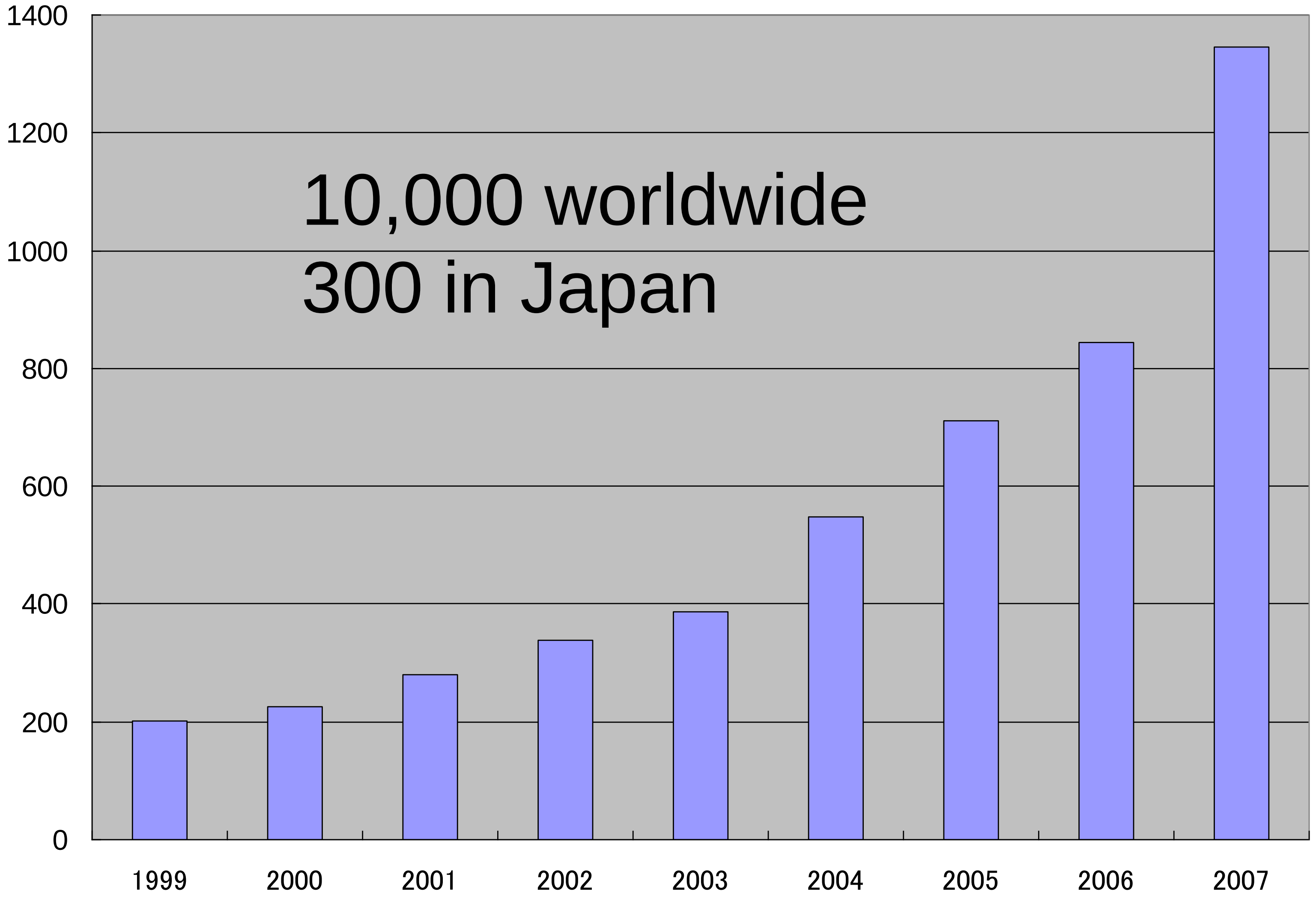
Many biological functions, from energy metabolism to antibiotic resistance, are carried out by biological pathways that require a number of cooperatively functioning genes. Hence, underlying mechanisms in the evolution of biological pathways are of particular interest. However, compared to the evolution of individual genes, which has been well studied, the evolution of biological pathways is far less understood. In this study, we used the abundant genome sequences available today and a novel algorithm we recently developed to trace the evolutionary history of prokaryotic metabolic pathways and to analyze how these pathways emerged. We found that the pathways have experienced significantly rapid acquisition, which would play a key role in eliminating the difficulty in holding genes during the course of pathway evolution. In addition, the emergence of novel pathways was suggested to have occurred more contemporaneously than expected across different phylogenetic clades. Based on these observations, we propose that novel pathway evolution can be facilitated by bidirectional horizontal gene transfers in prokaryotic communities. This simple model may approach the question of how biological pathways requiring a number of cooperatively functioning genes can be obtained and are the core event within the evolution of biological pathways in prokaryotes.

→

Automatic extraction

Molecular interaction
Disorder–gene relationships
Gene expression control
Intracellular location
Biological pathways
Protein functions
Function and concept hierarchies
Experimental conditions

Number of Databases (from *NAR*)



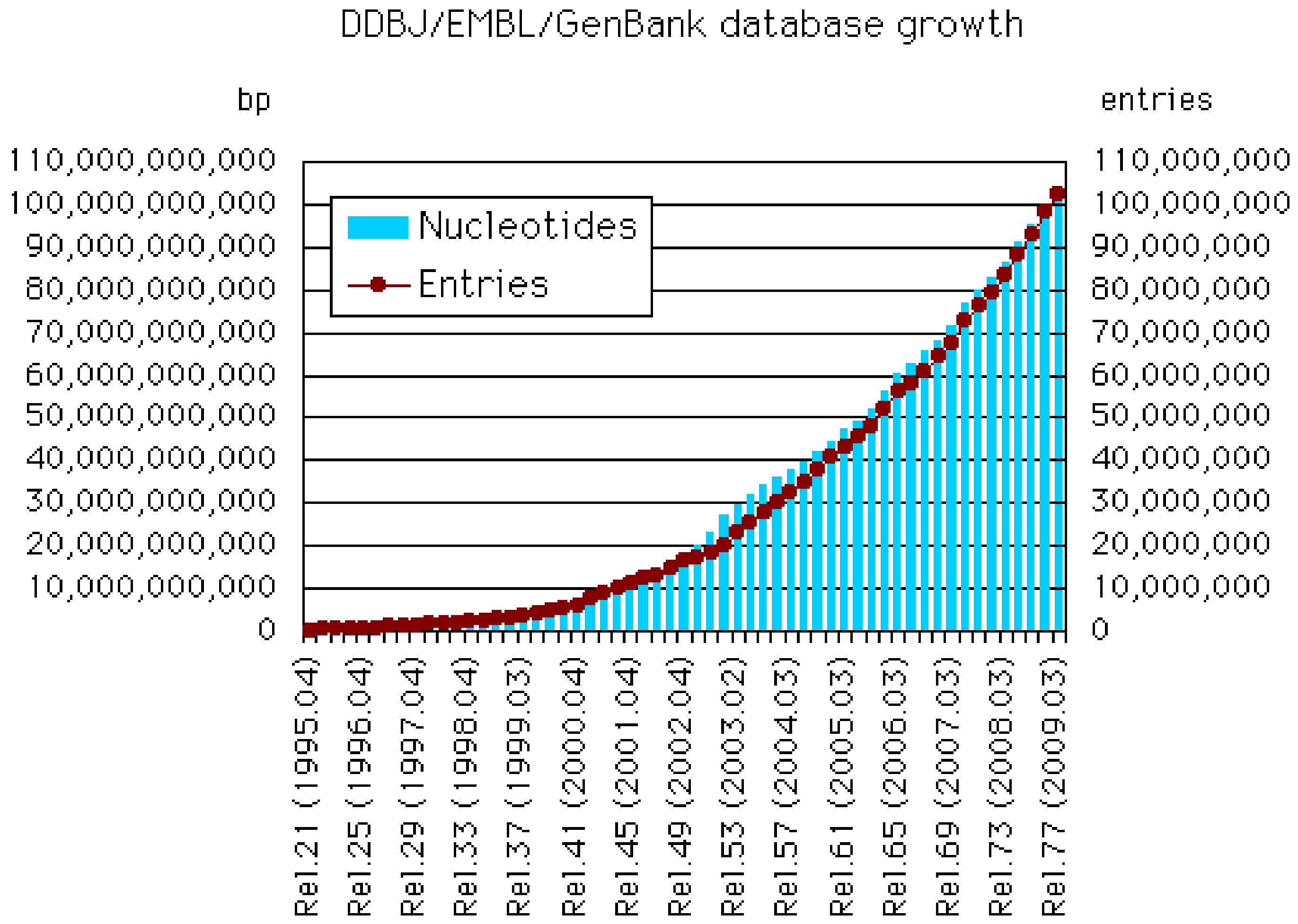
10,000 worldwide

300 in Japan

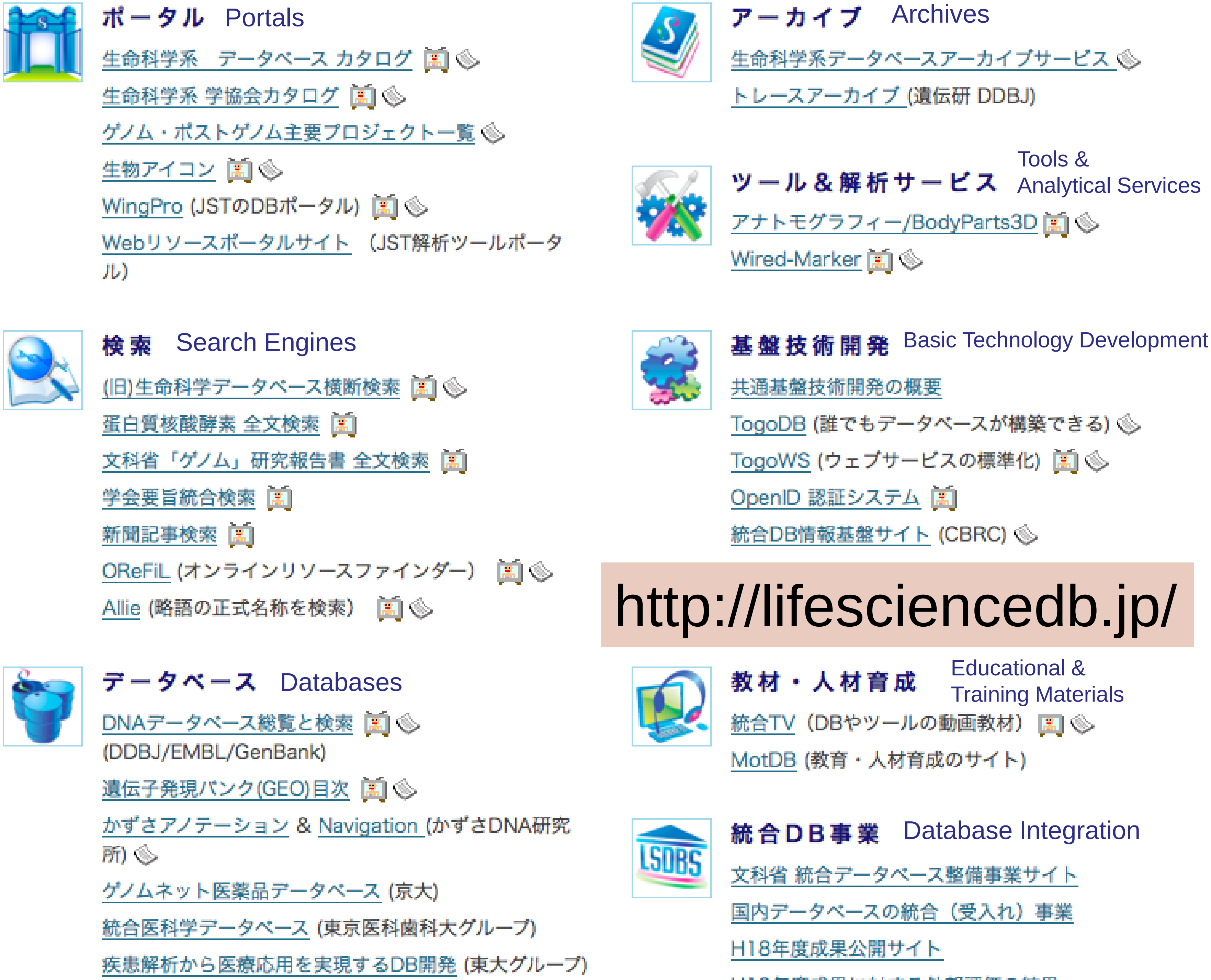
The Paradigm Shift in Research and the Importance of Sharing and Integrating Data and Knowledge

- Progress in genome research
- Development of high-throughput, high-precision instrumentation
- From hypothesis-driven studies to data-driven studies
- Corporatization, Bayh-Dole Act --> Data hoarding
- Explosion of data, explosion of knowledge, explosion of databases
- Subdivision of fields, fragmentation and segmentation of data and knowledge
- Understanding systems requires sharing and integration

Growth in DDBJ Nucleotides



‡ * Note : CON and TPA divisions are not counted in the Release statistic.



ポータル Portals

生命科学系　データベース カタログ

生命科学系 学協会カタログ

ゲノム・ポストゲノム主要プロジェクト一覧

生物アイコン

WingPro (JSTのDBポータル)

Webリソースポータルサイト （JST解析ツールポータル）

検索 Search Engines

(旧)生命科学データベース横断検索

蛋白質核酸酵素 全文検索

文科省「ゲノム」研究報告書 全文検索

学会要旨統合検索

新聞記事検索

OReFiL (オンラインリソースファインダー）

Allie (略語の正式名称を検索)

データベース Databases

DNAデータベース総覧と検索 (DDBJ/EMBL/GenBank)

遺伝子発現バンク(GEO)目次

かずさアノテーション & Navigation (かずさDNA研究所)

ゲノムネット医薬品データベース (京大)

統合医科学データベース (東京医科歯科大グループ)

疾患解析から医療応用を実現するDB開発 (東大グループ)

アーカイブ Archives

生命科学系データベースアーカイブサービス

トレースアーカイブ (遺伝研 DDBJ)

ツール＆解析サービス Tools & Analytical Services

アナトモグラフィー/BodyParts3D

Wired-Marker

基盤技術開発 Basic Technology Development

共通基盤技術開発の概要

TogoDB (誰でもデータベースが構築できる)

TogoWS (ウェブサービスの標準化)

OpenID 認証システム

統合DB情報基盤サイト (CBRC)

http://lifesciencedb.jp/

教材・人材育成 Educational & Training Materials

統合TV（DBやツールの動画教材）

MotDB (教育・人材育成のサイト)

統合DB事業 Database Integration

文科省 統合データベース整備事業サイト

国内データベースの統合（受入れ）事業

H18年度成果公開サイト

Summary

- The spread of genome studies and their applications
- Different approaches to understanding systems
- Importance of computers and databases
- Integration of many different fields
- Cancer, brain, nanotech, physics, complex systems sciences, chemistry, bioethics…
- Paradigm shift in research
- Analysis more important than generating data
- New theories and methodologies needed