BT 623 Research Methodology

Types of Data in Biological Research

Prof. Utpal Bora
Department of Biosciences and Bioengineering
Indian Institute of Technology Guwahati
Kamrup, Assam- 781039, India
Email: ubora@ iitg.ac.in

1. Sequences.

Sequence data, such as those associated with the DNA of various species, have grown enormously with the development of automated sequencing technology.

Perspectives

Anecdotal, Historical and Critical Commentaries on Genetics

Edited by James F. Crow and William F. Dove

The First Sequence: Fred Sanger and Insulin

Antony O. W. Stretton

Department of Zoology, University of Wisconsin, Madison, Wisconsin 53706

FRED Sanger is an amazingly modest man, and his own retrospective, written after he retired, a delightful prefatory chapter for the Annual Reviews of Biochemistry, is called "Sequences, sequences, and sequences" (SANGER 1988). In it he describes the paths that led to the successful methods he developed for the sequencing of proteins, then RNA, and then DNA. What a career!

Especially now, with the human genome largely finished, it is almost impossible to imagine a world without sequences of proteins and of nucleic acids. The fact that it has been only 50 years since Sanger showed that there into the Laboratory of Molecular Biology), took me on as a research student. Vernon had just shown that sickle-cell hemoglobin differed from normal hemoglobin by a single amino acid substitution, the first characterization of the molecular consequences of mutation on proteins (INGRAM 1956, 1957); earlier NEEL (1949) had shown that sickle-cell anemia is inherited as a Mendelian character, and PAULING *et al.* (1949) showed that sickle-cell hemoglobin differed electrophoretically from normal hemoglobin and coined the term "molecular disease." I had been trained as a chemist and knew nothing about proteins: I had heard Alex Todd's exciting lec-

FIGURE 1.—The structure of bovine insulin.

be isolated and identified. Fred showed that there were four N-terminal residues per 12K insulin molecule, two of which were glycine and two phenylalanine (SANGER 1945), suggesting that there were four polypeptide chains in the 12K molecule. Cysteine was present, so it was thought that the chains were held together by -S-S- bridges, and indeed after performic acid oxidation, which splits the -S-S- bridges, insulin could be fractionated by precipitation into an A fraction and a B fraction; the A fraction had N-terminal glycine, and the B fraction had phenylalanine (SANGER 1949). The two fractions had different amino acid compositions, and neither contained tryptophan. Later it became clear that the 12K molecule is a noncovalent dimer of the fundamental molecular unit (HARFENIST and CRAIG 1952), comprising one A chain and one B chain (see Figure 1).

The lack of tryptophan was particularly fortunate, because it degrades upon acid hydrolysis, and one of the most important methods Fred used to get at the structure of insulin, with great success, was partial acid hydrolysis. This article (SANGER 1949) was pivotal—it showed for the first time that at least some of the amino acids were in a unique sequence in insulin. Furthermore, the A and B fractions each yielded a unique sequence, suggesting that there were only two, not four, species of peptide chain in insulin—an A chain that contained about 20 amino acids and a B chain with about 30 amino acids—and he already had the sequence of over a quarter of the B chain! Even more important, this article showed that it should be possible in principle to determine the whole structure of each chain simply by extending the methods developed in this article—partial hydrolysis, fractionation of the products, end group analysis, and further partial hydrolysis of the longer products. In practice, it is the fractionation methods that are limiting, and the complexity of the mixture has to be controlled to match them. SANGER and TUPPY (1951a) did many experiments to approach a compromise between the ideal and the feasible, which meant concentrating on the later stages of the hydrolysis, where the average size of the peptides, and therefore also their number in the mixture, were relatively low.

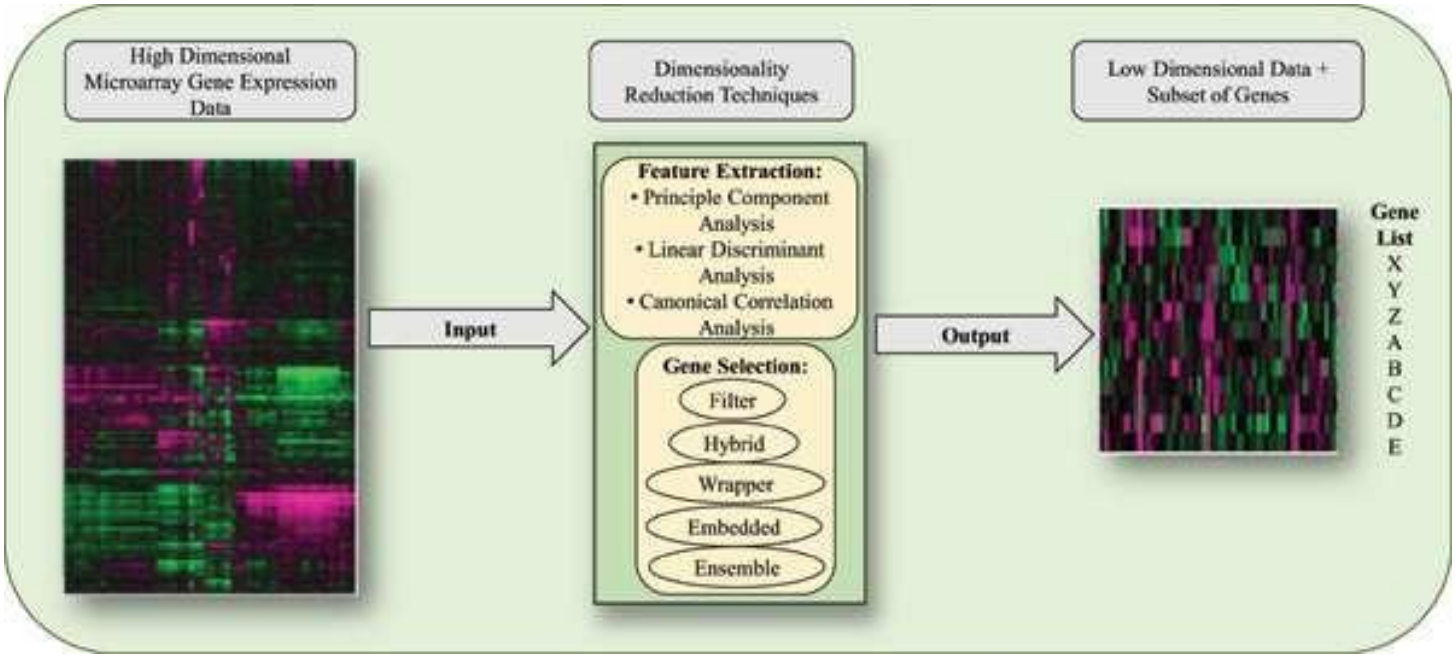
The complete sequence of the B chain: The B chain was tackled first (SANGER and TUPPY 1951a). Partial acid hydrolysis of the whole untagged chain yielded many more products to be separated, and the problem of separation of pure peptides from such a complex mixture was severe. They used several methods. To fractionate acidic peptides, they used batch absorption on ion

2. High-dimensional data.

Systems biology is highly dependent on comparing the behavior of various biological units. So data points that might be associated with the behavior of an individual unit must be collected for thousands or tens of thousands of comparable units.

Example, gene expression experiments can compare expression profiles of tens of thousands of genes.

The variation of expression profiles as a function of different experimental conditions (say 10^2 - 10^3) could yield 10^6 to 10^7 data points to be analyzed.

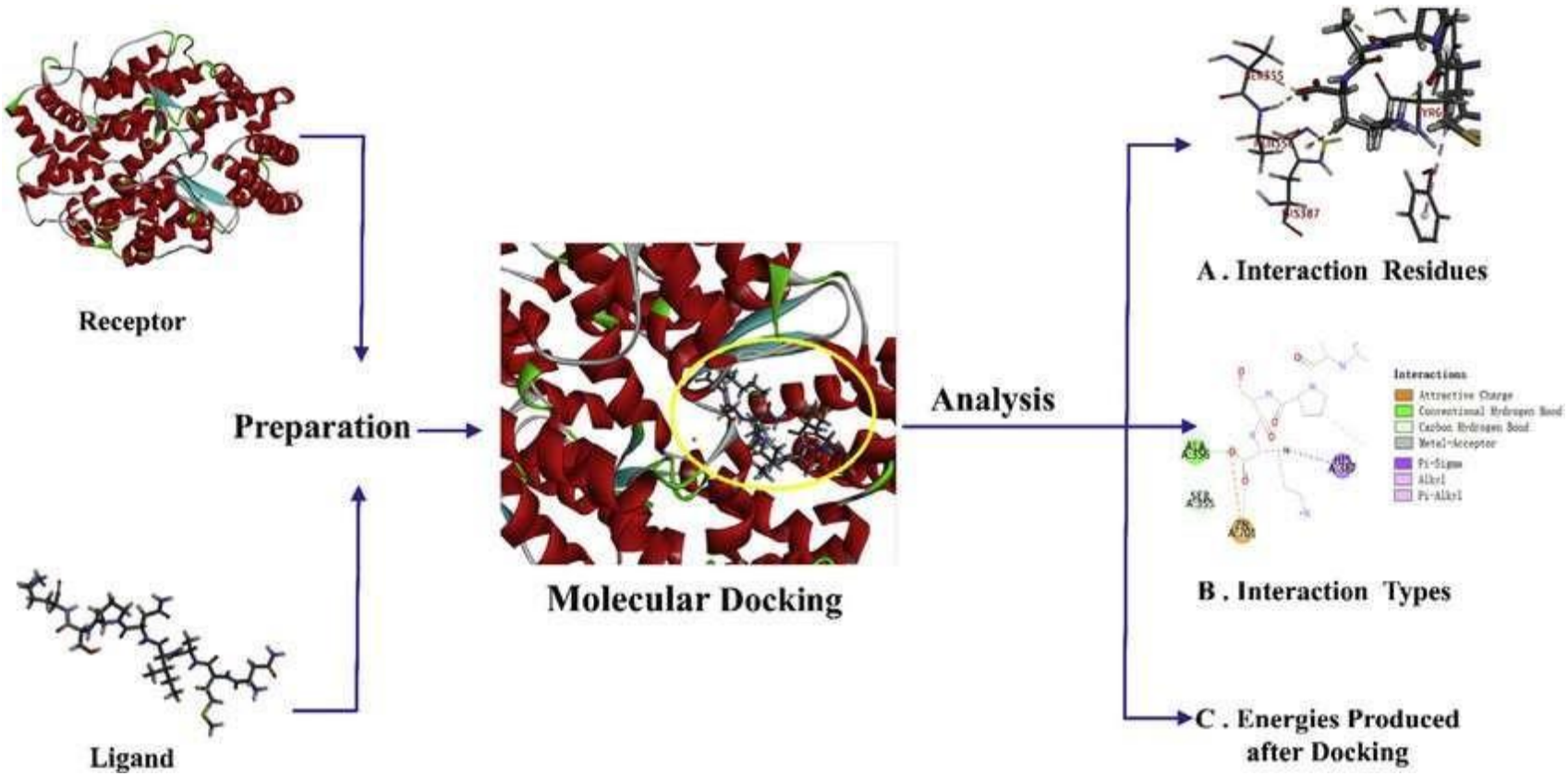

High Dimensional Microarray Gene Expression Data
Input
Dimensionality Reduction Techniques
Feature Extraction:
• Principle Component Analysis
• Linear Discriminant Analysis
• Canonical Correlation Analysis
Gene Selection:
Filter
Hybrid
Wrapper
Embedded
Ensemble
Output
Low Dimensional Data + Subset of Genes
Gene List
X
Y
Z
A
B
C
D
E

3. Geometric information.

A great deal of biological function depends on relative shape so molecular structure data are very important.

e.g., the “docking” behavior of molecules at a potential binding site depends on the three-dimensional configuration of the molecule and the site.

Graphs are one way of representing three-dimensional structure, but ball-and-stick models of protein backbones provide a more intuitive representation.



https://www.researchgate.net/figure/General-procedures-for-molecular-docking_fig3_324508414 [accessed 29 Sep, 2020]

4. Scalar and vector fields.

Scalar and vector field data are relevant to natural phenomena that vary continuously in space and time.

In biology, scalar and vector field properties are associated with

- chemical concentration and electric charge across the volume of a cell,
- current fluxes across the surface of a cell or through its volume, and
- chemical fluxes across cell membranes,

The data regarding charge, hydrophobicity, and other chemical properties that can be specified over the surface or within the volume of a molecule or a complex.

5. Patterns.

Within the genome are patterns that characterize biologically interesting entities.

e.g. the genome contains patterns associated with genes (i.e., sequences of particular genes) and with regulatory sequences (that determine the extent of a particular gene's expression).

Proteins are characterized by particular genomic sequences.

Patterns of sequence data can be represented as

- regular expressions,
- hidden Markov models (HMMs),
- stochastic context-free grammars (for RNA sequences),
- or other types of grammars.

Patterns are also important in the exploration of

- protein structure data,
- microarray data,
- pathway data,
- proteomics data, and
- metabolomics data.

6. Constraints.

Consistency within a database is critical if the data are to be trustworthy, and biological databases are no exception.

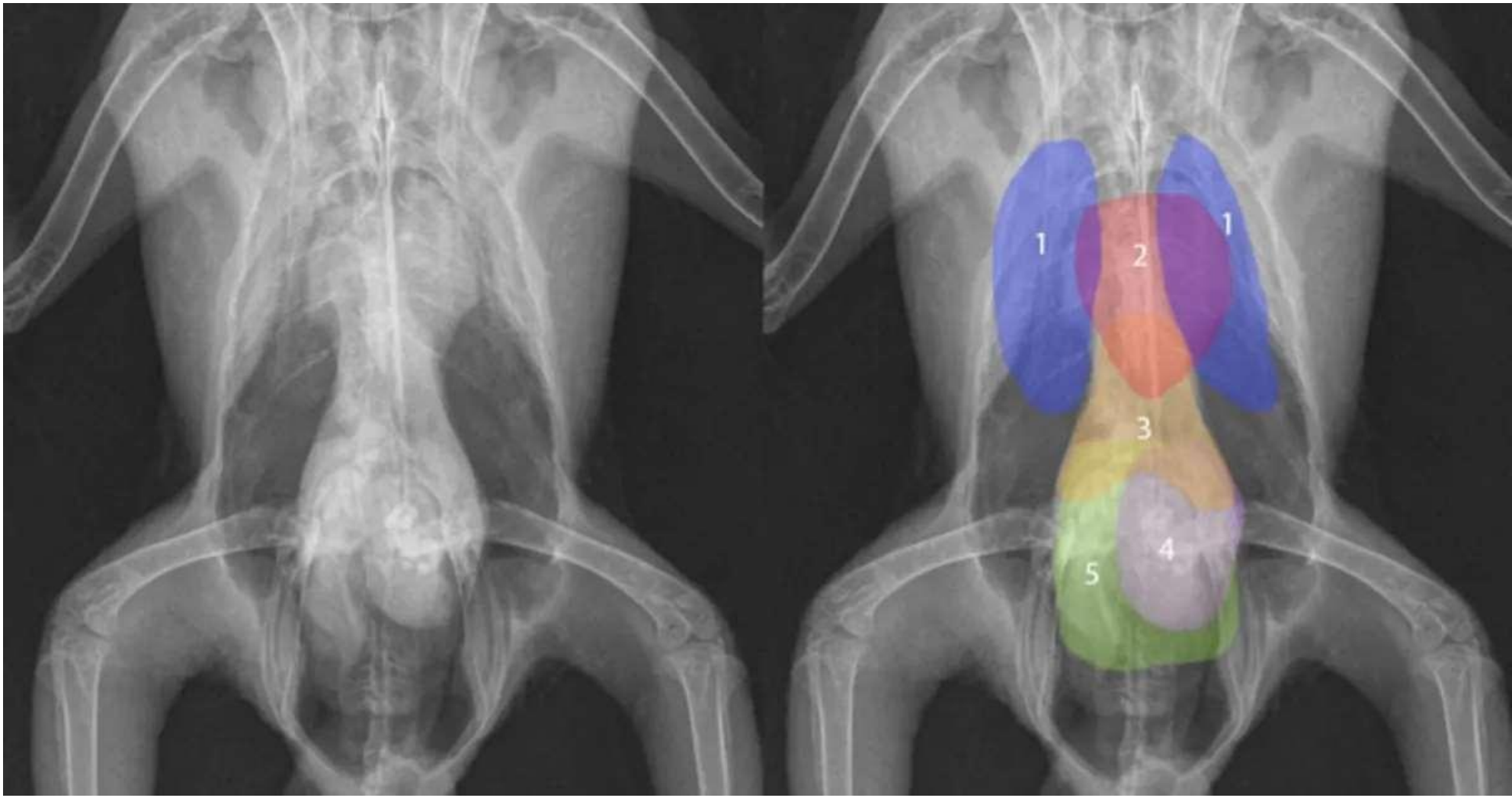
E.g. individual chemical reactions in a biological pathway must locally satisfy the conservation of mass for each element involved.

Reaction cycles in thermodynamic databases must satisfy global energy conservation constraints.

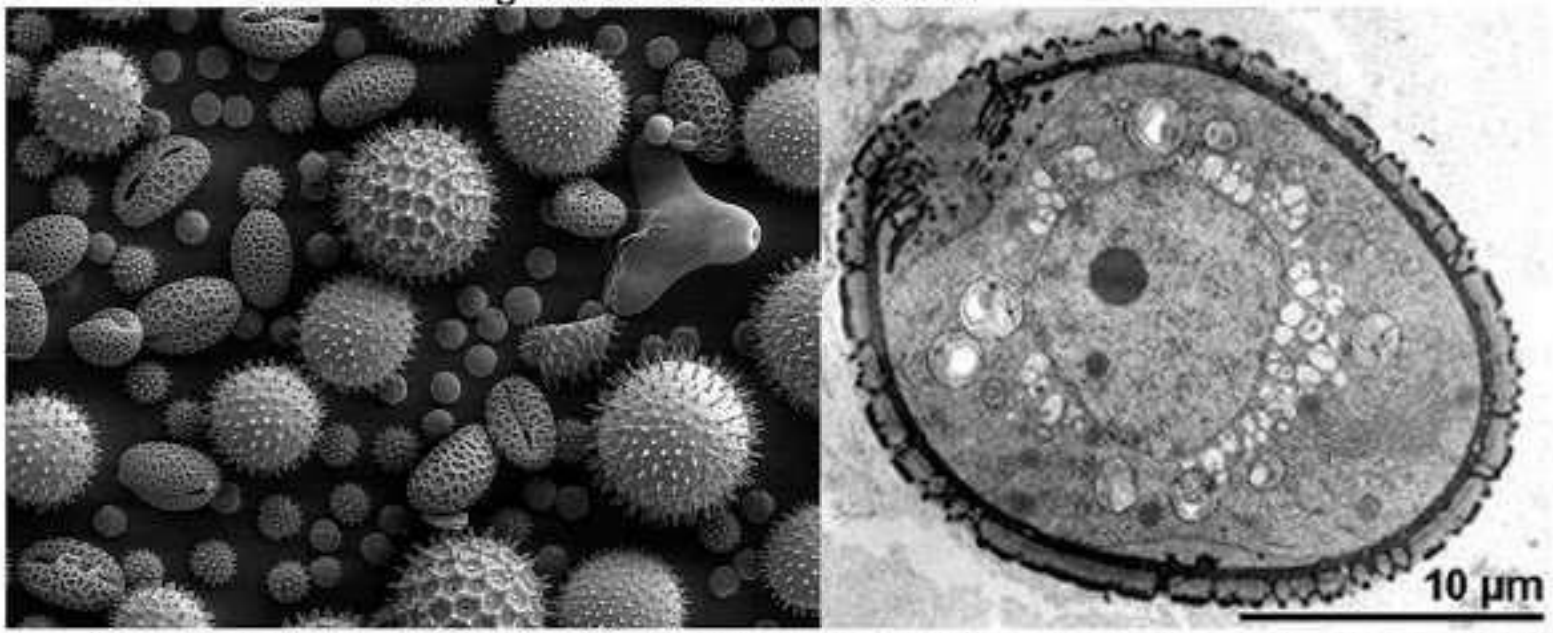
7. Images.

Imagery is an important component of biological research. Imagery may be either natural or artificial.

- Electron and optical microscopes are used to probe cellular and organ function.
- Radiographic images are used to highlight internal structure within organisms.
- Fluorescence is used to identify the expressions of genes.
- Cartoons are used to simplify and represent complex phenomena.
- Animations and movies are used to depict the operation of biological mechanisms over time and to provide insight and intuitive understanding that far exceeds what is available from textual descriptions or formal mathematical representations.


1
2
1
3
4
5

Pollen grain under SEM and TEM



Scanning Electron Microscope (SEM) vs Transmission Electron Microscope(TEM)

www.majordifferences.com

8. Spatial information.

Real biological entities, from cells to ecosystems, are not spatially homogeneous, and a great deal of interesting science can be found in understanding how one spatial region is different from another. Thus, spatial relationships must be captured in machine-readable form, and other biologically significant data must be overlaid on top of these relationships.

9. Models.

Computational models must be compared and evaluated. As the number of computational models grows, machine-readable data types that describe computational models—both the form and the parameters of the model—are necessary to facilitate comparison among models.

10. Prose.

The biological literature itself can be regarded as data to be exploited to find relationships that would otherwise go undiscovered. Biological prose is the basis for annotations, which can be regarded as a form of metadata. Annotations are critical for researchers seeking to assign meaning to biological data.

11. Declarative knowledge such as hypotheses and evidence.

As the complexity of various biological systems is unraveled, machine-readable representations of analytic and theoretical results as well as the underlying inferential chains that lead to various hypotheses will be necessary if relationships are to be uncovered in this enormous body of knowledge.

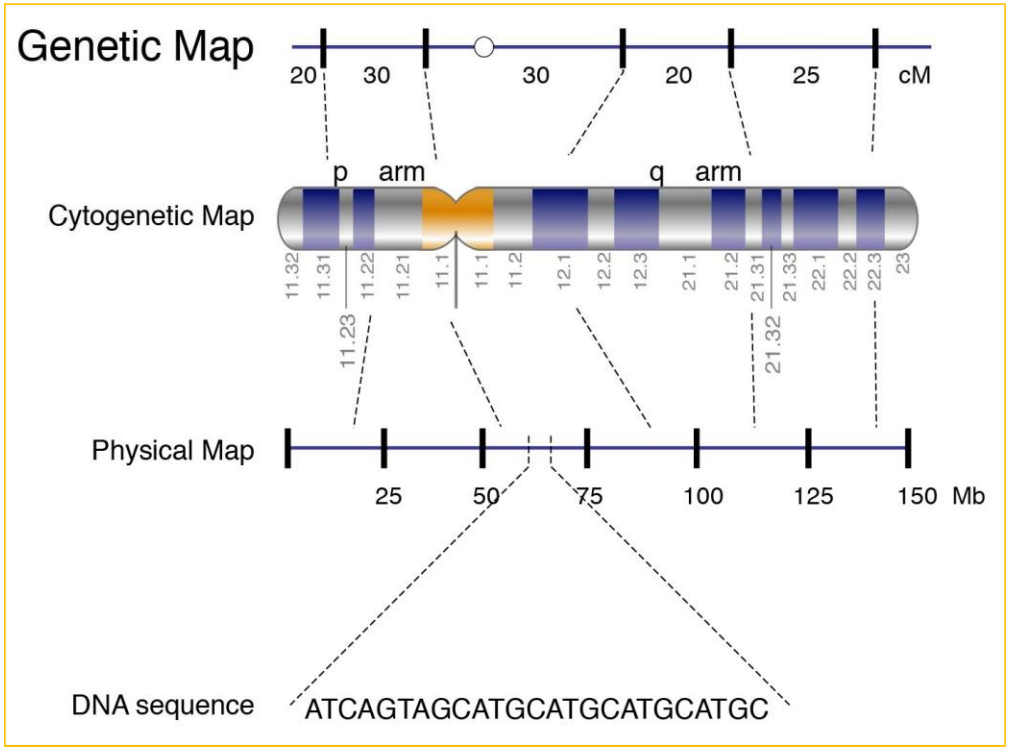
12. Graphs.

Biological data indicating relationships can be captured as graphs, as in the cases of pathway data, genetic maps, and structured taxonomies.

Even laboratory processes can be represented as workflow process model graphs and can be used to support formal representation for use in laboratory information management systems.

A genetic map is a type of chromosome map that shows the relative locations of genes and other important features.

Assignment:
Describe Cytogenetic and Physical Map



14. *Pathway Data*

In biological systems, **pathway data** refers to information about the biochemical routes or networks that represent the series of molecular events or interactions within cells.

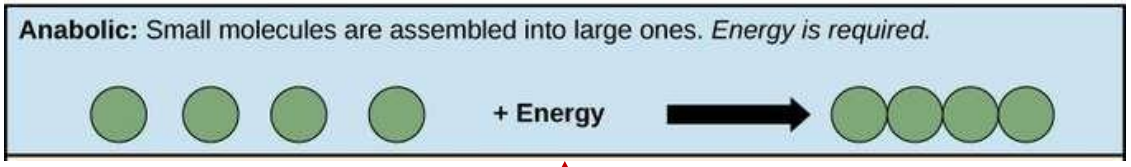
Pathways describe how cells carry out essential functions, such as metabolism, signaling, gene regulation, and cellular processes.

Common types of pathway data in biology include:

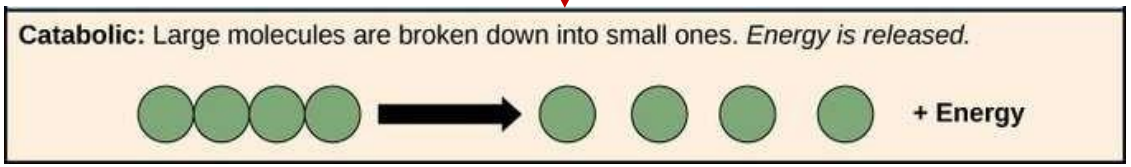
- Metabolic Pathways
- Signaling Pathways
- Gene Regulation Pathways
- Protein-Protein Interaction Pathways
- Disease Pathways

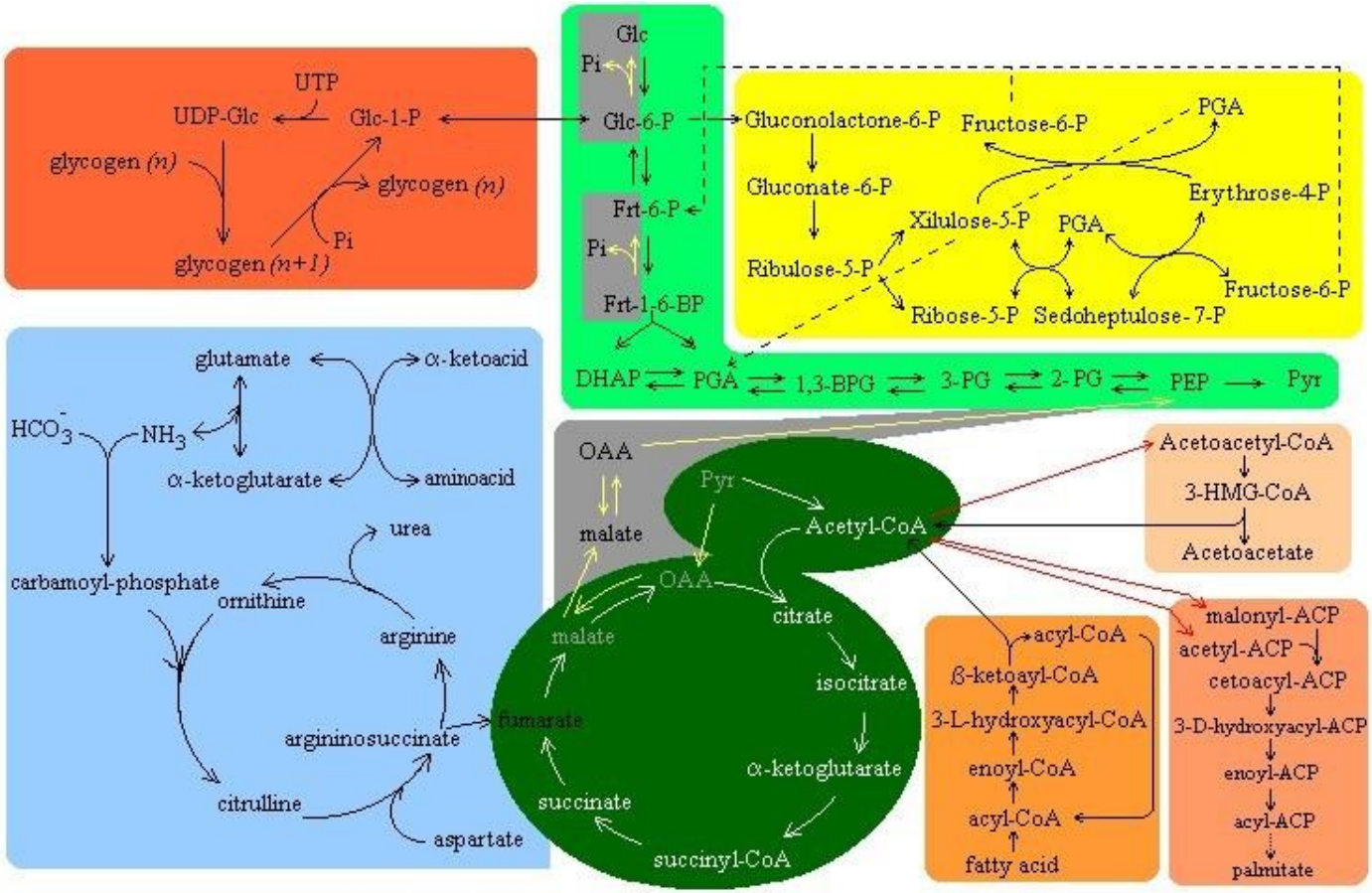
A **metabolic pathway** is a series of chemical reactions in a cell that build and breakdown molecules for cellular processes.

Anabolic pathways synthesize molecules and require energy.



Catabolic pathways break down molecules and produce energy.




UTP
UDP-Glc
Glc-1-P
glycogen (n)
glycogen (n)
Pi
glycogen (n+1)
Glc
Pi
Glc-6-P
Frt-6-P
Pi
Frt-1,6-BP
Gluconolactone-6-P
Fructose-6-P
PGA
Gluconate-6-P
Erythrose-4-P
Xilulose-5-P
PGA
Ribulose-5-P
Fructose-6-P
Ribose-5-P
Sedoheptulose-7-P
DHAP
PGA
1,3-BPG
3-PG
2-PG
PEP
Pyr
glutamate
α-ketoacid
HCO3-
NH3
α-ketoglutarate
aminoacid
urea
carbamoyl-phosphate
ornithine
arginine
argininosuccinate
fumarate
citrulline
aspartate
OAA
malate
Pyr
Acetyl-CoA
OAA
citrate
malate
isocitrate
α-ketoglutarate
fumarate
succinate
succinyl-CoA
Acetoacetyl-CoA
3-HMG-CoA
Acetoacetate
acyl-CoA
β-ketoayl-CoA
3-L-hydroxyacyl-CoA
enoyl-CoA
acyl-CoA
fatty acid
malonyl-ACP
acetyl-ACP
cetoacyl-ACP
3-D-hydroxyacyl-ACP
enoyl-ACP
acyl-ACP
palmitate

Assignment:

signaling pathways,
gene regulatory networks,
structured taxonomies

DATA ACCURACY AND CONSISTENCY

All laboratories must deal with instrument-dependent or protocol-dependent data inconsistencies. For example, measurements must be calibrated against known standards, but calibration methods and procedures may change over time, and data obtained under circumstances of heterogeneous calibration may well not be comparable to each other. Experiments done by multiple independent parties almost always result in inconsistencies in datasets.

Different experimental runs with different technicians and protocols in different labs inevitably produce data that are not entirely consistent with each other, and such inconsistencies have to be noted and reconciled. Also, the absolute number of data errors that must be reconciled—both within a single dataset and across datasets—increases with the size of the dataset. For such reasons, statistical data analysis becomes particularly important in analyzing data acquired via high-throughput techniques.

To illustrate these difficulties, consider the replication of microarray experiments. Experience with microarrays suggests that such replication can be quite difficult. In principle, a microarray experiment is simple. The raw output of a microarray experiment is a listing of fluorescent intensities associated with spots in an array; apart from complicating factors, the brightness of these spots is an indication of the expression level of the transcript associated with them.
On the other hand, the complicating factors are many, and in some cases ignoring these factors can render one's interpretation of microarray data completely irrelevant.

Consider the impact of the following:

•**Background effects**, which are by definition contributions to spot intensity that do not originate with the biological material being examined. For example, an empty microarray might result in some background level of fluorescence and even some variation in background level across the entire surface of the array.

•**Noise dependent on expression levels of the sample**. For example, Tu et al. found that hybridization noise is strongly dependent on expression level, and in particular the hybridization noise is mostly Poisson-like for high expression levels but more complex at low expression levels.

•**Differential binding strengths for different probe-target combinations**. The brightness of a spot is determined by the amount of target present at a probe site and the strength of the binding between probe and target. Held et al. found that the strength of binding is affected by the free energy of hybridization, which is itself a function of the specific sequence involved at the site, and they developed a model to account for this finding.

•**Lack of correlation between mRNA levels and protein levels.** The most mature microarray technology measures mRNA levels, while the quantity of interest is often protein level. However, in some cases of interest, the correlation is small even if overall correlations are moderate. One reason for small correlations is likely to be the fact that some proteins are regulated after translation, as noted in Ideker et al.

•**Lack of uniformity in the underlying glass surface of a microarray slide**. Lee et al. found that the specific location of a given probe on the surface affected the expression level recorded.

Discuss the pros and cons microaaray Data

thankyou