Semester -III

18BTP305B **SYSTEM BIOLOGY** **4H – 4C**

Total hours/week: **L: 4 T:0 P:0** **Marks:** Internal:**40** External:**60** Total: **100**

Course Objectives:

Course Outcomes:

Scope: The Systems Biology is designed primarily for highly motivated postgraduate students interested in interdisciplinary studies in life sciences, behavioral sciences, and engineering and computer sciences.

Objective: Apply a network biology analysis approach to a wide rage of molecular biology problems. Describe the main high-throughput experimental methods used for generating protein-protein interaction data.

UNIT - I

Introduction to Systems Biology: Introduction to Systems Biology. Need for System Analysis in Biology. Basic Concepts in System Biology: Component vs System, Links and Functional States, Links to Networks, Hierarchical Organization in Biology.
systems, scales, static/dynamic, approaches, limitations, reductionism; central dogma; mathematical models; computational analysis; statistics of prokaryotes and eukaryotes.

UNIT - II

Metabolic Networks and Models in System Biology: Basic Features of Metabolic Networks. Reconstruction Methods of Metabolic Networks. Models as Dynamical Systems. SYN1, SYN3 and molecular simulation, Parameter Problem. Meanings of Robustness.

UNIT - III

Systems Biology Databases KEGG (Kyoto Encyclopedia of Genes and Genomes). BRENDA (BRaunschweig ENzyme DAtabase). BioSilico. EMP (Embden-Meyerhof-Parnas). MetaCyc and AraCyc. SABIO-RK (System for the Analysis of Biochemical Pathways - Reaction Kinetics). BioModels.

UNIT - IV

Tools for System Biology: Cell Designer. Ali Baba. Cell Profiler. JDesigner. Bio-SPICE (Biological Simulation Program for Intra and Inter Cellular Evaluation). SBML (Systems Biology Markup Language). SBGN (Systems Biology Graphical Notation). SBML-SAT (SBML based Sensitivity Analysis Tool).

UNIT - V

Premises & Promises of Systems Biology: Premise of Systems Biology. Promise of Systems Biology. Challenges of Systems Biology. Applications of Systems Biology.

References

Bernhard Palsson, O. (2006). *Systems Biology: Properties of Reconstructed Networks.* New York: Cambridge University Press.

Björn Junker, H., Falk Schreiber. (2008). *Analysis of Biological Networks.* New Jersey: John Wiley & Sons, Inc.

Huma Lodhi, M., & Stephen Muggleton, H. (2010) *Elements of Computational Systems Biology*. New Jersey: John Wiley & Sons, Inc.

Cánovas, M., Iborra, J.L., & Manjón, A. (2006). *Understanding and Exploiting Systems Biology in Biomedicine and Bioprocesses.* Spain: CajaMurcia Foundation.

Brown, T. A. (2006). *Genomes* (2nd ed.). UK: BIOS Scientific Publishers, Ltd.

Sensen, C.W. (2002). *Essentials of Genomics and Bioinformatics*, Wiley-VCH.
Pennington, S.R. & Dunn, M.J. (2002). *Proteomics.* New Delhi: Viva Books Pvt. Ltd.

http://www.systemsbiology.org

http://www.systems-biology.org

KARPAGAM ACADEMY OF HIGHER EDUCATION

DEPARTMENT OF BIOTECHNOLOGY

II M.Sc., BIOTECHNOLOGY – SEMESTER 3 LECTURE PLAN –SYSTEM BIOLOGY

18BTP305C

S.No	Lecture Duration (hr)	Topic to be covered	Support materials
		UNIT I	
1.	1	Introduction to Systems Biology	T2 Pg 3-5
2.	1	Need for System Analysis in Biology	T3 Pg 3-7
3.	1	Basic Concepts in System Biology, Component vs. System	
4.	1	Links and Functional States, Links to Network	T2 Pg 17-18
5.	1	Hierarchical Organization in Biology. Systems, scales, static/dynamic, approaches, limitations, reductionism	T1 Pg 215
6.	1	Central dogma; mathematical models	T3 Pg 174-175;161-165
7.	1	Computational analysis	T3 Pg 11-13
8.	1	Statistics of prokaryotes and eukaryotes.	T2 Pg 40-45
9.	1	Revision	
10.	1	Test	
		UNIT II	
11.	1	Introduction to metabolic networks and models in system biology	T2 Pg 19-20
12.	1	Basic Features of Metabolic Networks	T2 Pg 19-20
13.	1	Reconstruction Methods of Metabolic Networks	T2 Pg 22-23
14.	1	Dynamic systems – as models	T1 Pg 256
15.	1	SYN1, SYN3 and Introduction	T1 Pg 300-305
16.	1	SYN1, SYN3 and molecular simulation	T1 Pg 300-305
17.	1	Parameter Problem	T3 Pg 31-34
18.	1	Meanings of Robustness	T2 Pg 25-26
19.	1	Revision	
20.	1	Test	

UNIT III			
21	1	System Biology Database – Introduction	T1 Pg76-77
22	1	KEGG (Kyoto Encyclopedia of Genes and Genomes)	T1 Pg76-77
23	1	BRENDA (Braunschweig Enzyme Database	T1 Pg 148-150 T3Pg 17,174-175
24	1	BioSilico.	J1 Pg 169-179
25	1	EMP (Embden-MeyerhofParnas)	J1 Pg 169-179
26	1	MetaCyc and AraCyc	T3 Pg 189, 207
27	1	SABIO-RK (System for the Analysis of Biochemical Pathways – Reaction Kinetics).	T1 Pg 149-150
28	1	BioModels	T2 Pg 511; W1
29	1	Revision	
30	1	Test	
UNIT IV			
31	1	Tools for system Biology- Introduction	T2 Pg 422
32	1	Cell Designer. Ali Baba	T1 Pg 66,73 113
33	1	Cell Profiler. Jdesigner	T1 Pg 114-115
34	1	Bio-SPICE (Biological Simulation Program for Intra and Inter Cellular Evaluation)	T1 Pg 64 T3 Pg 4,116,117-118,123,207
35	1	SBML (Systems Biology Markup Language)	T1 Pg 70, T3 Pg 4,15,19,20,27,28,30,84,103
36	1	SBGN (Systems Biology Graphical Notation)	
37	1	SBML-SAT (SBML based Sensitivity Analysis Tool).	T2 Pg 395;W2
38	1	Revision	
39	1	Test	
UNIT V			
40	1	Premises of system biology	T3 Pg 1-35
41	1	Promises of system biology	T3 Pg 2-3
42	1	Challenges of System biology	T3 Pg 129
43	1	Applications of system biology	T3 Pg 31-34
44	1	Revision (Theory)	
45	1	Revision (MCQ)	
46	1	Advancement in System Biology	T3 Pg 9-11
47	1	Recent trends in system Biology	T3 Pg 12-13
48	1	Test	

Reference Books

1. T1. Axel Kowald, Christoph Wierling, Edda klip Hans Lehrach & Ralf Herwig (2008), System Biology, Wiley Blackwell.
2. T2: Sangdun Choi (2007), Introduction to System Biology, Human Press Publications.
3. T3: Andres Kriete, Roland Elis (2006), Computational System Biology, Elsevier Academy Press. Harcourt academic press.
4. W1: http:www.ebi.ac.uk/biomodels
5. W2: http:www.http://sbgn.org
6. J1: Kitano. H (2003), A Graphical notation for Biological Network: Biosilico: 1: 169-179.

UNIT-I

SYLLABUS

Unit – I

Introduction to Systems Biology: Introduction to Systems Biology. Need for System Analysis in Biology, Basic Concepts in System Biology: Component vs. System, Links and Functional States, Links to Networks, Hierarchical Organization in Biology. Systems, scales, static/dynamic, approaches, limitations, reductionism; central dogma; mathematical models; computational analysis; statistics of prokaryotes and eukaryotes.

Introduction

Suddenly, systems biology is everywhere. What is it? How did it arise? The driving force for its growth is high-throughput (HT) technologies that allow us to enumerate biological components on a large scale. The delineation of the chemical interactions of these components gives rise to reconstructed biochemical reaction networks that underlie various cellular functions. Systems biology is thus not necessarily focused on the components themselves, but on the nature of the links that connect them and the functional states of the networks that result from the assembly of all such links. The stoichiometric matrix represents such links mathematically based on the underlying chemistry, and the properties of this matrix are key to determining the functional states of the biochemical reaction networks that it represents.

The Need for Systems Analysis in Biology

Biological parts lists During the latter half of the 20th century, biology was strongly influenced by reductionist approaches that focused on the generation of information about individual cellular components, their chemical composition, and often their biological functions. Over the past decade, this process has been greatly accelerated with the emergence of genomics. We now have entire DNA sequences for a growing number of organisms, and we are continually delineating their gene portfolios. Although functional assignment to these genes is presently incomplete, we can expect that we will eventually have assigned and verified function for the majority of genes on selected genomes. Extrapolation between genomes will then most likely accelerate the definition of what amounts to a "catalog" of cellular components in a large number of organisms. Expression array and proteomic technologies

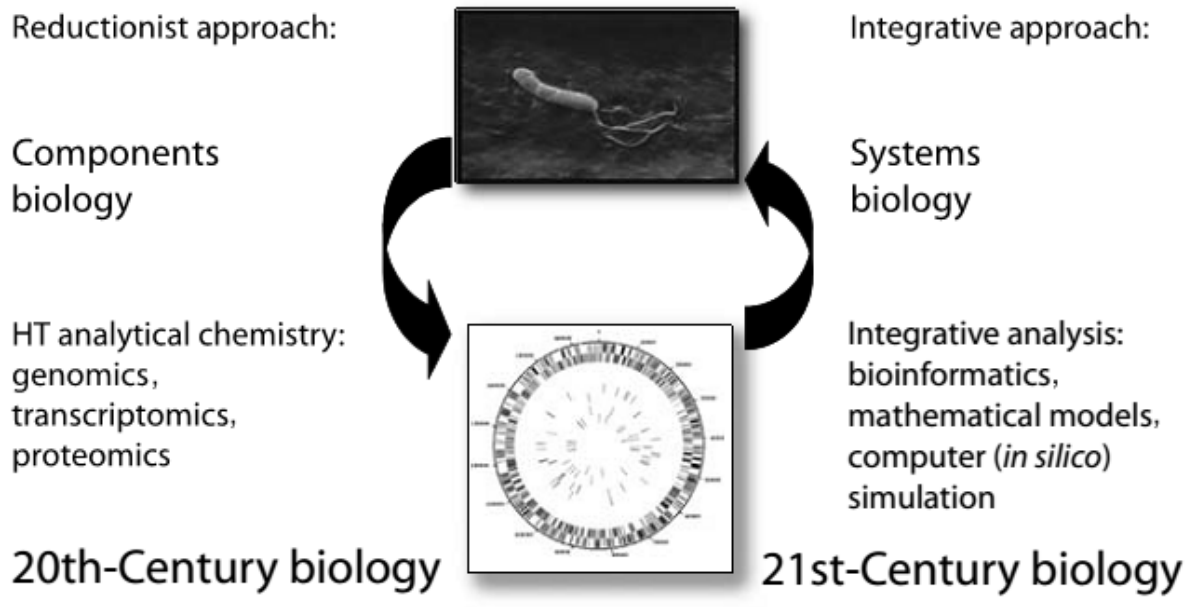

Figure 1.1: Illustration of a paradigm shift in cell and molecular biology from component to systems analysis. Redrawn from [152].

give us the capability to determine when a cell uses particular genes, and when it does not (left side in Figure 1.1). At the beginning of the 21st century, this process was unfolding at a rapid rate, driving a fundamental paradigm shift in biology.

Beyond bioinformatics

The advent of high-throughput experimental technologies is forcing biologists to view cells as systems, rather than focusing their attention on individual cellular components. Not only are high-throughput technologies forcing the systems point of view, but they also enable us to study cells as systems. What do we do with this developing list of cellular components and their properties? As informative as they are, these lists only give us basic information about the molecules that make up cells, their individual chemical properties, and when cells choose to use their components. How do we now arrive at the biological properties and behaviors that arise from these detailed lists of chemical components? It is now generally accepted that the integrative analysis of the function of multiple gene products has become a critical issue for the future development of biology. Such integrative analysis will rely on bioinformatics and methods for systems analysis (right side of Figure 1.1). It is thus likely that over the coming years and decades biological sciences will be increasingly focused on the systems properties of cellular and tissue functions. These are the properties that arise from the whole and represent biological properties. These properties are sometimes referred to as “emergent” properties since they emerge from the whole and are not properties of the individual parts.

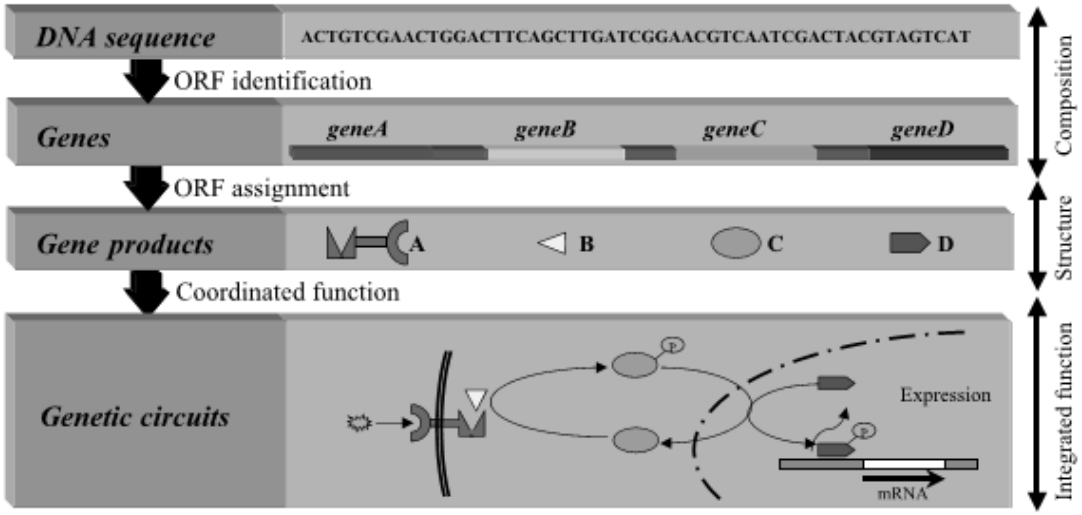

Figure 1.2: Genetic circuits. From sequence, to genes, to gene product function, to multicomponent cellular functions. Prepared by Christophe Schilling.

Genetic circuits

The relationship between genetics and cellular functions is hierarchical and involves many layers, some of which are illustrated in Figure 1.2. Gene sequences allow for the identification of open reading frames (ORFs). The base pair sequence of the ORFs in turn allows for the functional assignment of the defined gene. Although not always unambiguous, such assignments are being carried out with increasing accuracy, due to our expanding biological databases. Sequence is important and so is the functional assignment of ORFs. However, the interrelatedness of the genes may prove to be even more important. Establishing these relations and studying their systemic characteristics is now necessary. Cellular functions rely on the coordinated action of the products from multiple genes. Such coordinated function of multiple gene products can be viewed as a "genetic circuit" (some synonyms that are commonly used are "cellular wiring diagrams" and modules). The term genetic circuit is used here to designate a collection of different gene products that together are required to execute a particular cellular function. The

functions of such genetic circuits are diverse, including DNA replication, translation, the conversion of glucose to pyruvate, laying down the basic body plan of multicellular organisms, and cell motion. It is likely that we will view cellular functions within this framework and the physiological functions of cells and organisms as the coordinated or integrated functions of multiple genetic circuits. Consequently, we will need to develop the conceptual framework within which to describe and analyze these circuits. Not all the properties of genetic circuits are clear at present, but some important ones are summarized in Table 1.1. For many of these characteristics, it is also clear what methodology is needed to describe and analyze

Table 1.1: Some of the characteristics of genetic circuits and the analysis methods required.

Characteristic	Analysis method
They are complex	Bioinformatics
They are autonomous	Control theory
They are robust	System science
They function to execute a physicochemical process	Transport and kinetic theory
They have "creative functions"	Bifurcation analysis
They are conserved, but can adjust	Evolutionary dynamics

them. Genetic circuits tend to have many components; they are complex. From the standpoint of system science, they are "robust," i.e., in many, but not all cases, one can remove their components without compromising their overall function.

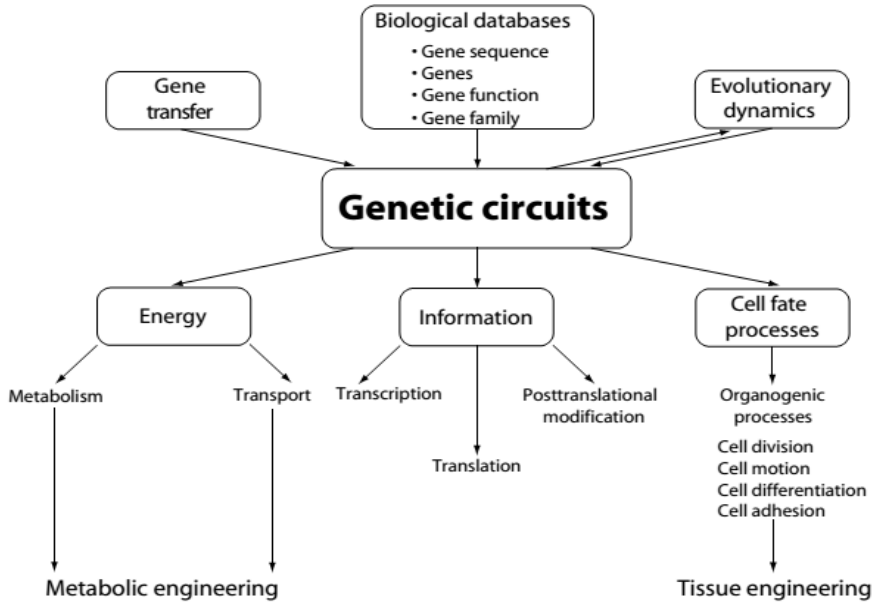

Figure 1.3: Coarse-grained classification of the types of genetic circuits that are found on genomes. Some major categories are indicated, in particular to indicate that some underlie the important metabolic and tissue engineering applications of cell and molecular biology. Prepared by Christophe Schilling.

Accepting the concept of a gene circuit seems straightforward. However, the implications of this acceptance are quite profound. We will view bioinformatics as a way to establish, classify, and cross-species correlate genetic circuits. The beginning of such classification is illustrated in Figure 1.3. Metabolism, information processing, and cellular fate processes represent some of the major categories of genetic circuits. Considerable unity in biology is likely to result in conceptualizing biological functions as genetic circuits. From this standpoint, gene therapy may no longer be viewed as replacing a "bad" gene, but instead fixing a "malfunctioning" genetic circuit. Evolution may be viewed as the "tuning" or "honing" of circuits to improve performance and chances of survival. Classifying organisms based on the types of genetic circuits they possess may lead to "genomic taxonomy." Ex vivo "evolutionary" procedures for designing genetic circuit performance are emerging [99, 258]. Understanding the function of genetic circuits will become fundamental to applied biology, in fields as diverse as metabolic

engineering and tissue engineering. The concept of a genetic circuit as a multicomponent functional entity (either in time or space, or both) is an important paradigm in systems biology. It will be a fundamental component in our treatment of the relationship between genetics and physiology. the genotype–phenotype relationship. Individual genetic circuits do not operate in isolation, but in the context of other genetic circuits. The assembly of all such circuits found on a genome produces cellular and organism functions and leads to hierarchical decomposition of complex cellular functions. Thus, the need for genome-scale analysis arises. This need in turn leads to viewing the genome as the "system."

Basic Concepts in Systems Biology

In the early 1960s, there was a bifurcation of emphasis in biology. Molecular biology had arrived, providing a growing understanding of DNA, protein, and other chemical components of cells. A science was emerging that had rigor in terms of analytical chemistry and controlled experimentation, and relevance to biochemical and genetic functions of cells and occasionally to their phenotypes. Holistic emphasis in biology, which had primarily been practiced through physiology, faded into the background as it is much more difficult to state hypotheses, do controlled experiments, or execute the scientific process for the behavior of systems and networks in biology. However, as outlined in the introductory chapter, this situation has now changed. We now have technology that allows for the detailed enumeration of biological components, enabling us to study cells and complex biological processes as systems. As a consequence, systems biology has arisen as a new field. This new field does not yet have a well-defined and articulated conceptual basis. In this chapter, we will attempt to collect some of the

key issues that represent to the conceptual foundations of systems biology. Its content is not intended to be, and cannot be, complete but rather represents an attempt to initiate this process.

Components vs. Systems

Biological components all have a finite turnover time. Most metabolites turn over within a minute in a cell, mRNA molecules typically have 2-hour half-lives in human cells [256], 3% of the extracellular matrix in cardiac muscle is turned over daily, and so forth. So a cell that you observe today, compared with the same cell yesterday, may only contain a small fraction of the same molecules. Similarly, cells have finite lifetimes. The cellularity of the human bone marrow turns over every 2–3 days, the renewal rate of

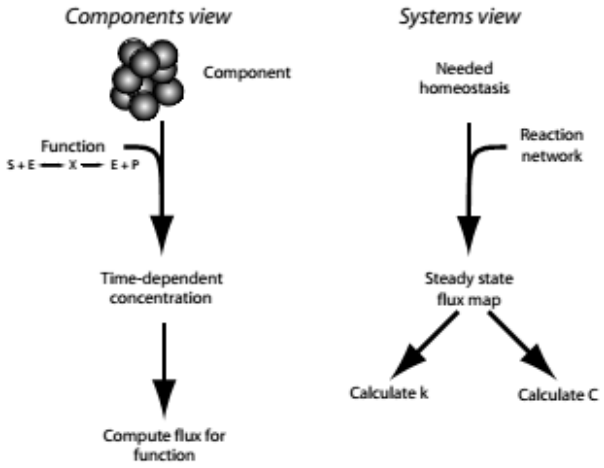

Figure 2.1: A contrast between the components view (left) and the systems view (right).

skin is of the order of 5 days to a couple of weeks, the lining of the gut epithelium has a turnover time of about 5–7 days, and slower tissues like the liver turn over their cellularity approximately once a year. Therefore, most of the cells that are contained in an individual today were not there

a few years ago. However, we consider the individual to be the same, albeit older. Likewise, we consider one cell to be the same a week later, even if most of its chemical components have turned over. Components come and go, therefore a key feature of living systems is how their components are connected together. The interconnections between cells and cellular components define the essence of a living process. The difference between the components view of life is different from the systems view in many subtle ways. Here, we try to illustrate this difference by just one example (see Figure 2.1).

On the left side of Figure 2.1 we see the components point of view of the function of a gene product. When we are looking at one gene product, in this case an enzyme carrying out its function, we study this component by placing it in a beaker with its substrates and then observe the time-dependent disappearance of a substrate and the appearance of a product. The component that we are studying is the centerpiece of this experiment, and it is responsible for concentration changing in a time-dependent manner. The right side illustrates a systems viewpoint of a biochemical network. It is not so much the components themselves and their state that matters, contrary to the components view, but it is the state of the whole system that counts. Any biological network will have a nominal state, which we recognize as a homeostatic state. Thus, the fluxes that reflect the interactions among the components to form the state of the network are dominant variables, and the concentrations of the individual components are "subordinate quantities." The concentrations of the network components are determined first by the flux map, or the state of the network, and then by the kinetic properties of the links in the network.

Links and Functional States

Two key issues arise from the earlier considerations. The first deals with the nature of the links between components in a biological network, and the second deals with the functional states and the properties of a network that a set of links form.

Links

Links between molecular components are basically given by chemical reactions or associations between chemical components. These links are therefore characterized and constrained by basic chemical rules. In tissue biology, the nature of links between cells is more complicated and often related to higher-order chemistry. We note that a T-cell receptor, for instance, forms a complicated structure in the membrane of a cell and the properties of that structure, and how compatible it is with the complimentary features of another cell determines whether there is communication or links between these cells. Since we are focused on the characteristics of biochemical networks, we will further discuss the chemical nature of links in molecular biology. The prototypical transformations in living systems at the molecular level are bilinear. This association involves two compounds coming together to either be chemically transformed through the breakage and formation of covalent bonds, as is typical of metabolic reactions or macromolecular synthesis,

X + Y X − Y covalent bonds

or two molecules associated together to form a complex that may be held together by hydrogen bonds and/or other physical association forces to form a complex that has a different functionality than individual components,

X + Y X : Y association of molecules

Such association, for instance, could designate the binding of a transcription factor to DNA to form an activated site to which an activated polymerase binds. Such bilinear association between two molecules might also involve the binding of an allosteric regulator to an allosteric enzyme that induces a conformational change in the enzyme.

Chemical transformations have certain key properties:

1. Stoichiometric. The stoichiometric of chemical reactions is fixed and is described by integral numbers counting the molecules that react and that form as a consequence of the chemical reaction. Thus, stoichiometric basically represents "digital information." Chemical transformations are constrained by elemental and charge balancing, as well as other features. Stoichiometric is invariant between organisms for the same reactions and does not change with pressure, temperature, or other conditions. Stoichiometric gives the primary topological properties of a biochemical reaction network.

2. Relative rates. All reactions inside a cell are governed by thermodynamics. The relative rate of reactions, forward and reverse, is therefore fixed by basic thermodynamic properties. Unlike stoichiometric, thermodynamic properties do change with physicochemical conditions such as pressure and temperature. The thermodynamic properties of associations between macromolecules can be changed by altering the sequence of a protein or the base-pair sequence of a DNA binding site. The thermodynamics of transformation between small molecules in cells are fixed but condition dependent.

3. Absolute rates. In contrast to stoichiometric and thermodynamics, the absolute rates of chemical reactions inside cells are highly manipulable. Highly evolved enzymes are very specific in catalyzing particular chemical transformations. Cells can thus extensively manipulate the rates of reactions through changes in their DNA sequence. Therefore, links cannot just form between any two cellular components. The links that are formed are constrained by the nature of covalent bonds that are possible and by the thermodynamic nature of interacting macromolecular surfaces. All of these are subject to the basic rules of chemistry and thermodynamics. The absolute rates are key biological design variables because they can evolve from a very low rate determined by the mass action kinetics based on collision frequencies to a very high and specific reaction rate determined by appropriately evolved enzyme properties. Enzymes evolve to bring molecules into particular orientation to control the rate of appropriately oriented collisions between two molecules that lead to a chemical reaction (see Figure 2.2).

Functional states

Once all the links in a network have been identified and described, its functional states can be determined. We can study the topological properties of a network, but these properties give is only limited information about the actual functional state of a network. The functional states of biological reaction networks are constrained by the physicochemical nature of the intracellular environment (see Figure 2.3). There is a highly developed spatiotemporal organization that orients the biological components and determines the transient nature of the interactions. Interestingly, cells are in a near crystalline state. The protein density in cytoplasm and mitochondria is very close to the protein density in a protein crystal. There are some other notable higher-order properties of biological networks, which will not be detailed here, which

include self-assembly of components to spontaneously form a functioning network, the selection that seems to be at work at all levels in biology, and the interesting notion of a self in biology, namely, is a component a part of a network or not?

Links to Networks

Chemical reactions link components together to form a network. Although we can specify the chemical properties of links in biological networks, it is the way in which a multitude of such links form networks that determines phenotypic functions. Most biochemical reactions are bilinear. Bilinearity gives the networks a hyper graph property that is topologically nonlinear. The biochemical consequence of this is that biochemical reaction networks form a tangle of cycles [186] where different chemical properties and moieties are being transferred throughout the network from one carrier to the next. Perhaps the most familiar of such transformations is the movement of high-energy phosphate bonds between metabolites and proteins. ATP is the primary carrier of such high-energy bonds, and, for instance, a phosphate group is tied to glucose to form glucose-6-phosphate as the first step in glycolysis. The same feature is found in signaling networks whose components are in phosphorylated or dephosphorylated states. Other properties being transferred between molecules are redox potential, 1 carbon units, 2 carbon units, ammonia groups, and so on. This makes biochemical reaction networks highly interwoven. One interesting feature of biochemical networks as they grow in size is the fact that because of combinatorics, the number of possible functional states that they can take can grow faster than the number of components in a network. This proliferation in the number of functional states seems to occur past some (a relatively low number) components that come together to form a network. Therefore, the number of phenotypic functions derivable from a genome does not linearly scale

with the gene number contained in that genome. For instance, the human genome may have only 50% more genes than the genome of *Caenorhabditis elegans*, a small worm, but nevertheless human beings display much more complicated phenotypes and in greater variety. Thus, in general, it is hard to correlate organism complexity and functions to the number of genes that the organism's genome contains.

The fundamental property of biochemical networks of having many possible functional states leads to the possibility of having the same network carry out many functions and displaying many different phenotypic behaviors. An organism does not fully exploit or use all such possible functional states. Many possible states will be useless to the organism in its struggle for survival. Therefore, a limited subset of these functional states needs to be selected and expressed by cells. We are becoming increasingly familiar with the regulatory mechanisms that carry out the selection of functional states. We are unraveling the very complicated transcriptional regulatory networks in single-celled organisms and the signaling networks that coordinate the function of multicellular organisms. As we will discuss in Chapter 16, complex biological reaction networks will have equivalent functional states; that is, there are identical overall functional states that differ in the ways in which they use the underlying links in the network. Some of the key features of biological networks that distinguish them from other networks need to be accounted for in the analysis of their systemic properties. The first basic feature of biological networks is that they evolve; they change with time. They are time variant. Principally, such changes occur through the kinetic properties of the links in the network and the changing of the available or active links in the network at any given point in time. The number of available links can be manipulated by regulation of gene expression, by horizontal gene transfer, and by other mechanisms. The second feature that has to be taken into account is the fact that they have a

sense of purpose. The fundamental purpose is survival. However, in complicated organisms that fundamentally comprise many networks, some will have goals that are subtasks to the overall goal of survival. For instance, the goal of adipocytes would be to collect and store fat if, in their environment, there is an abundance of energy resources. The goal of the mitochondrion, being the powerhouse of the cell, seems to be to maximize ATP production from available resources. Therefore, the study of objectives, that is, purpose, of biochemical reaction networks becomes a relevant and perhaps a central issue. Thus, linking many biological components together forms a network. This network can have many functional states from which a subset is selected. Links, network topology, and the sense of purpose can all change with time or environmental conditions. It is important to be cognizant of the fact that biochemical reaction networks have to operate in the crowded interior of a cell (see Figure 2.3). Thus, the network view of the biological process has to be considered in the context of the three-dimensional physical arrangement of such networks. These considerations may limit the usefulness of analogies with other man-made networks, such as electrical circuits.

Hierarchical organization in biology

Many facets of cellular function and properties are organized hierarchically. The spatial organization of the DNA is shown in Figure 2.6A. The linear dimension of the E. coli genome is about 1 mm, while the length of the cell is of the order of 1 m, a 1000-fold difference. The bacterial genome is thus "folded" a thousand times, in a hierarchically organized fashion. Biochemical reaction networks can be similarly decomposed (Figure 2.6B). Reactions group together into coordinated units that may be co localized in space, or even compartmentalized. Many such coordinated units can form a larger organized unit, and so forth. The constraints that

apply to the lower levels of organization by necessity will constrain the subsequent higher level functions. This upward application of constraints necessitates a bottom-up approach to the analysis of complex biological phenomena. Go ¨del's completeness theorem in mathematics that showed an axiomatic approach to proving mathematical theorems could not prove that all properties of a system may in a general sense apply to biology. If so, we cannot construct all higher level functions from the elementary operations alone. Thus, observations and analyses of system level functions will be needed to complement the bottom-up approach. Therefore, bottom-up and top-down approaches are complementary to the analysis of the hierarchical nature of complex biological phenomena.

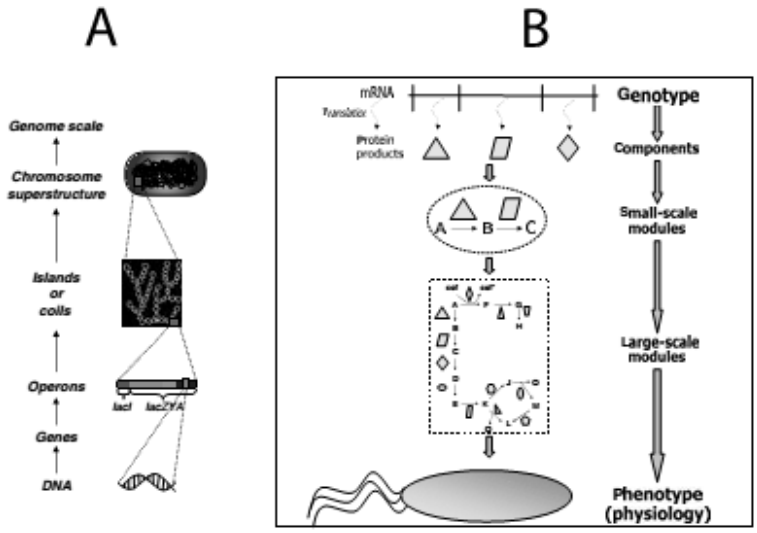

Figure 2.6: Hierarchical organization in cells: (a) bacterial genome and (b) network topology and function. Prepared by Timothy Allen (A) and Jason Papin (B).

The successive adoption of cellular functions over evolution is illustrated in Figure 2.7. The basic biochemistry of cellular processes and the maintenance and expression of the information on the DNA molecule evolved early. This basic set of processes is still found in most organisms today. The genetic code is essentially universal and most proteins are made up of about 20 amino

acids. These are basic constraints under which all subsequent cellular processes must operate. The genetic code cannot be predicted from basic theory or physics [39] but is consistent with the basic laws of physics and chemistry. Once picked, it is essentially fixed over evolution. Similarly, most modern proteins are made up of a limited number of motifs, and the basic circuits that lay out the body plan are remarkably conserved. Thus, the constraints set at a lower level of biological hierarchy confine higher levels of organization but may not explain or predict the more complex functions. Evolution is a "tinkerer" that combines the elements at hand together in new and unpredictable ways. The first "wave" in Figure 2.7 is close to the underlying chemical principles and will thus naturally represent a focus of this text.

UNIT-II

SYLLABUS

Unit – II

Metabolic Networks and Models in System Biology: Basic Features of Metabolic Networks. Reconstruction Methods of Metabolic Networks. Models as Dynamical Systems. SYN1, SYN3 and molecular simulation, Parameter Problem. Meanings of Robustness.

Metabolic Networks

The function of cells is based on complex networks of interacting chemical reactions carefully organized in space and time. These biochemical reaction networks produce observable cellular functions. Network reconstruction is the process of identifying all the reactions that comprise a network. The reconstruction process for metabolic networks has been developed and implemented for a number of organisms. The main features of metabolic network reconstruction are described in this chapter. We briefly review the key properties of metabolic networks and introduce the hierarchical thinking that goes into the interpretation of complex network functions. Further details can be found in authoritative sources. As discussed at the end of this chapter, a true genome-scale reconstruction of cellular functions necessitates accounting for all cellular networks simultaneously. Such a comprehensive network reconstruction has yet to be established; therefore, in this chapter, we focus on metabolism and address the reconstruction of transcriptional regulatory and signaling networks in the following two chapters.

Basic Features

Intermediary metabolism can be viewed as a chemical "engine" that converts available raw materials into energy as well as the building blocks needed to produce biological structures, maintain cells, and carry out various cellular functions. This chemical engine is highly dynamic, obeys the laws of physics and chemistry, and is thus limited by various physicochemical constraints. It also has an elaborate regulatory structure that allows it to respond to a variety of external perturbations. Metabolic imbalance is involved in major human diseases, such as diabetes, obesity, cancer, and heart disease. Metabolism comprises two types of chemical transformations: catabolic pathways that break down various substrates into common metabolites and anabolic pathways that collectively synthesize amino acids, fatty acids, nucleic acids, and other needed building blocks. During these processes, an intricate exchange of various chemical

groups and reduction oxidation (redox) potentials takes place through a set of carrier molecules (see Table 3.1). These carrier molecules and the properties that they transfer thus tie the metabolic network tightly together. Intermediary metabolism can be described at several levels of complexity (Figure 3.1).

Table 3.1: Key chemical groups in metabolism and their carriers.

Group carried in activated form	Carrier molecule
Phosphoryl	ATP, GTP
Electrons	NADH, NADPH, $FADH_2$, $FMNH_2$
One carbon unit	Tetrahydrofolate
Methyl	*S*-Adenosylmethionine
Acyl (two carbons)	Coenzyme A, lipoamine
Aldehyde	Thiamine pyrophosphate
Carbon dioxide	Biotin
Nucleotides	Nucleoside triphosphates

Hierarchy in function of metabolic networks

Genome-scale reconstructions of metabolic networks contain hundreds of metabolites and sometimes over a thousand reactions (see Table 3.6). The functions of such networks are hard for the human mind to comprehend. We thus need mathematical models for the study of their properties and simulation of their function. However, as pointed out in Section 1.2, we can think of network properties in a hierarchical fashion to simplify the conceptualization of network functions. Such hierarchy can be based on manmade concepts, as discussed later, or can be the result of a nonbiased mathematical analysis of the stoichiometric matrix (see Chapter 9). In what follows, we briefly describe the traditional view of the hierarchical decomposition of the functions of metabolic networks (see Figure 3.1).

Level 1: Cellular inputs and outputs. Overall, intermediary metabolism comprises the enzymatic reactions pertaining to the transformation of substrate molecules into the essential building blocks of macromolecules and other vital products for growth and maintenance. A coarse- grained description of the overall activity of metabolism thus involves substrates as inputs and biomass and metabolic by-products as outputs. For industrial fermentation processes, a description of cells at this level has sufficed for many purposes. The description comprises a simple set of coupled mass and energy balances, with various empirically determined "yield" coefficients that describe partitioning of the consumed substrate. Growth kinetics is given in terms of simple phenomenological models such as the Monod growth model. Models of this type are useful for a limited set of specific conditions. The yield coefficients are not constants; they change with the physiological state of the cell.

Level 2: Sectors. A bit finer grained look at intermediary metabolism reveals that it can be divided into two basic sectors (see Figure 3.2). Catabolism carries out the degradation of substrates via a series of converging pathways that lead to a set of 11 metabolites of central importance, called the biosynthetic precursors. Anabolism is a set of diverging pathways that originate from these central metabolites to form monomers or building blocks for macromolecular biosynthesis. Genetically engineered bacteria used for bioprocessing, for instance, can be described at this level of complexity since it is appropriate for assessing host- plasmid interactions.

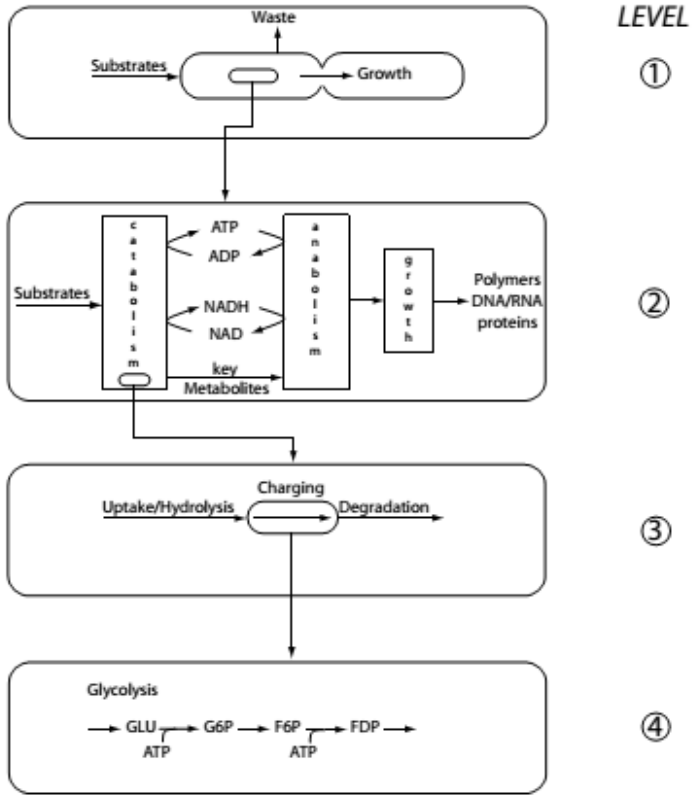

Figure 3.1: Four-level functional decomposition of metabolism. Level 1: whole cell; level 2: metabolic sectors; level 3: pathways; and level 4: individual reactions.

Level 3: Pathways. A still finer resolution reveals a situation in which pathways, and segments thereof, serve a definite role. For instance, catabolism of the major classes of biomolecules follows the same pattern; first, substrates are picked up by the cell, hydrolyzed if necessary, activated by a cofactor, and then degraded to yield energy and other properties stored on the carrier molecules. At this level of description, the essential features of metabolism begin to depend on basic chemical principles such as stoichiometric structure and kinetic regulation. Key metabolic pools, such as the energy charge, dominate the description, and key regulatory enzymes influence the motion of these pools and how mass and energy is distributed among them. There is currently much interest in the pathway level characterization of reconstructed biochemical reaction networks.

Level 4: Individual reactions. At the finest level of description one considers all the biochemical transformations that take place in a cell. Available high-throughput data, as discussed in Chapter 1, allows us to generate the information needed to describe cells at this resolution. It is at this level where this book is focused. We can now reconstruct genome-scale stoichiometric matrices of organisms and study them. The dimensions of these matrices are on the order of hundreds of metabolites and sometimes over a thousand chemical reactions, reflecting the complexity of a fully functional metabolic network.

Biochemical transformations fall into a few major categories. Some examples include transamination, phosphorylation, isomerization, dehydration, and dismutation. Thus, there are chemical “rules” that dictate what kind of links can exist in metabolic networks. As described later, biochemists have devised nomenclature that classifies these types of transformations and an Enzyme Commission (E.C.) number is associated with each enzymatically catalyzed

metabolic reaction. Furthermore, there are thermodynamic restrictions associated with these transformations that dictate the energetic feasibility of a reaction and its equilibrium state. Thus, even though metabolic networks may appear complex, there are underlying physicochemical restrictions on their topological structure and network states.

Reconstruction Methods

Defining the reaction list

The reconstruction of a genome-scale metabolic network relies on assembling various sources of information about all the biochemical reactions in the network. A variety of data sources can be used to synthesize a list of chemical reactions that form an organism's metabolic network (see Figure 3.3). The principal data sources are (roughly in the order of reliability) as follows:

1. Biochemistry. The strongest evidence for the presence of a metabolic reaction is found if an enzyme has been isolated directly from the organism and its function demonstrated. Extensive data is often available for model organisms, such as Escherichia coli and yeast but may be fragmented for organisms that have been sparingly studied.

2. Genomics. Functional assignments to open reading frames (ORFs), based on DNA sequence homology, may be used as a strong evidence for the presence of a reaction in an organism. Functional assignments can also be achieved from the genome location of an ORF and the cluster of genes that are found in its neighborhood. Genome annotations are subject to revision and updates.

3. Physiology and indirect information. Physiological evidence, such as the known ability of the cell to produce an amino acid in vivo, may lead us to include reactions which "fill in the

pathway" to produce that amino acid. This process is called gap analysis. Other physiological information is often useful in diagnosing the function of a reconstructed networks.

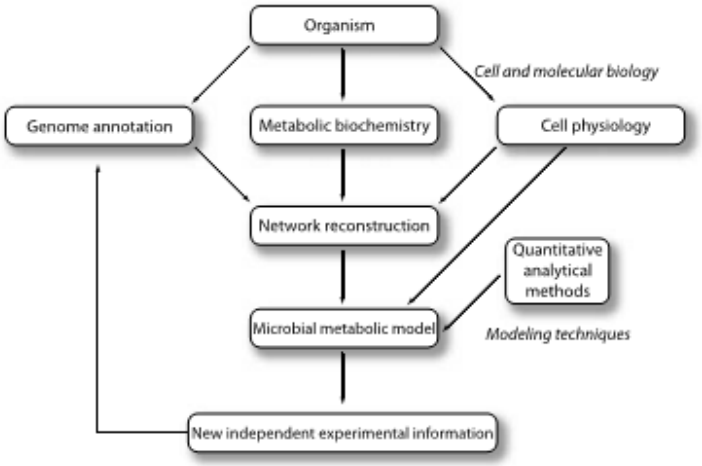

Figure 3.3: A schematic of the overall process of genome-scale metabolic network reconstruction (GENRE) and subsequent model formulation. Modified from [37].

4. *In silico* modeling data. Modeling and simulation studies often lead to the inclusion of metabolic reactions in the reconstruction. A network needs to be able to simulate cell behavior *in silico*. For instance, the metabolic network must be able to produce or take up all of the necessary components of the cellular biomass. One needs to add the reactions necessary to fulfill the biomass requirements if they are not present. Such reactions are referred to as "inferred reactions."

All the reactions identified by these various means then combine to produce a genome-scale metabolic reconstruction for the organism of interest. Normally, the reconstruction process starts with the annotated DNA sequence and thus the reconstruction is "genome-scale" since it will contain all the information that is found on the genome that relates to the organism's metabolism. This set of reactions comprises a genome-scale metabolic model when combined with quantitative analytical methods, which enable us to analyze, interpret, and predict integrated

network functions (see Figure 3.3). Some of these mathematical methods that are scalable to the genome-scale networks are described later in the text.

Table 3.2: Methods for annotating genomes [108, 148].

ORF identification	"Traditional" annotation methods	New annotation methods
Stop codons	Experimental (direct)	Protein–protein interactions
GLIMMER	Sequence homology	Correlated mRNA expression levels
Genscan		Phylogenetic profile
		Protein fusion clustering
		Gene neighbors (operon clustering)
		Automation

Clearly, the confidence level in the various sources differs, but one can use quantitative scale to rank-order the reliability of the source. One such quantitative scheme proposed is biochemical data, genetic data, genomic data, physiological data, and modeling data. One is never fully sure about the presence of a reaction until the biochemical data has been obtained, although sequence homology that meets certain criteria is often taken as sufficient evidence for a true functional assignment.

Genome annotation

Since few organisms have extensive biochemical information available, reconstruction relies heavily on an annotated genome sequence. ORFs are identified on the genomic sequence, then assigned a function. This process can be done through experimental methods (gene cloning and expression or gene knockout) or more commonly by comparing its sequence homology to genes of known function in other organisms. *In silico* annotation methods typically lead to functional assignment of 40–70% of identified ORFs on a freshly sequenced microbial genome. New and improved methods continue to be developed for genome annotation. For example, functions of

gene products may be inferred from protein–protein interactions, transcriptomics, phylogenetic profiles, protein fusion, and operon clustering (see Table 3.2). It should be emphasized that every gene annotation based on *insilico* methods is hypothetical, and such annotation is subject to revision until the gene has been cloned, expressed, and the function of the gene product directly evaluated. The automation of network reconstruction from annotated sequence has been attempted [108]. To produce high-quality, well-curated reconstructions, one has to manually verify all the components and links in a network, since there are often subtle differences between even related organisms. There are many Web resources available for this purpose (Table 3.3).

Publicly available sources of sequence data. There are several publicly available databases that contain genomic data (Table 3.3). The Comprehensive Microbial Resource (CMR) provides tools for the analysis of 63 annotated genome sequences, both individually and collectively. The Institute for Genomic Research (TIGR) updates and maintains this site. Another database that maintains many microbial genomes is the Genomes On-Line Database (GOLD) site. Not all of the information on the site is publicly available. The developers of GOLD have been active in automating the reconstruction of metabolic networks using pathway templates.

Biochemical data

Direct biochemical information is the most reliable source for the presence of a reaction in an organism. Biochemical data also gives stoichiometric and whether or not a reaction is reversible. For example, the enzyme that catalyzes the conversion of D-glucose to D-glucose-6-phosphate, as ATP is converted to ADP, is called glucokinase. The gene that encodes this enzyme is commonly called glk, and the E.C. number that corresponds to this reaction is 2.7.1.2. The structure of the Human β-cell glucokinase is shown at the top of Figure 3.5 (found in the Protein

Data Bank). Collections of biochemical data on an organism's metabolism is often found in review articles and more recently in whole volumes that are focused on the biology of a single organism.

Enzyme commission numbers

E.C. numbers are used to systematically characterize enzymatic reactions (http://www.chem.qmul.ac.uk/iubmb/enzyme/). They have been established to unambiguously classify reactions, which is needed because so many enzymes have ambiguous and duplicate names across organisms (see Table 3.4). For instance, try going to the E.C. Web site and searching first for succinate dehydrogenase (sdh), and then for fumarate reductase (fr). Both of these enzymes catalyze the same reaction, but in opposite directions. Some biochemists find that frd or sdh may be reversible at times. As a result, when you type in succinate dehydrogenase you will find that it is often used to indicate either reaction. A classification scheme similar to the E.C. system is being developed for transport reactions [26]. Unfortunately, there is no similar system for genes, which have the same problem of ambiguous and duplicate names. Thus, the curation of gene annotation information for a reconstruction can be quite laborious.

Protein databases

Swiss-Prot (http://us.expasy.org/sprot/) is a very useful source for examining particular protein or reaction assignments in detail and is considered a " gold standard" for biochemical information because it is so well-curated. It contains literature references, sequences, functional assignments, and other useful information, all specific to the organism being examined. If one is not sure about the presence of a protein in an organism but a page is found for it on Swiss-Prot,

he or she can be fairly sure that the protein has been characterized and that literature references are available. TrEMBL contains new entries to Swiss-Prot that have not yet been curated.

Gene–protein–reaction (GPR) associations

When associating genes to reactions, and vice versa, it is important to remember that not all genes have a one-to-one relationship with their corresponding enzymes or metabolic reactions. Many genes may encode subunits of a protein which catalyze one reaction. One example is the fumarate reductase. There are four subunits, frdA, frdB, frdC, and frdD, without which the enzyme (a protein complex) will not be able to catalyze the reaction. Conversely, there are genes that encode so-called promiscuous enzymes that can catalyze several different reactions, such as transketolase I in the pentose phosphate pathway. Such reactions typically involve similar chemical transformations of structurally related molecules. These examples highlight the need to keep track of associations between genes, proteins, and reactions. Examples of different types of GPR associations are shown in Figure 3.6, where the top level is the gene locus, the second level is the translated peptide, the third level is the functional protein, and the bottom level is the reaction. Many genes may encode subunits of a protein, or multiple proteins might come together to form an enzyme complex. Subunits (e.g., sdhABCD and gapC1C2) and enzyme complexes (e.g., xylFGH) are connected to reactions with "&" associations, meaning that all have to be expressed for the reaction to occur. For sdhABCD, the "&" is shown above the enzyme level indicating that all of these gene products are needed for the functional enzyme. With xylFGH, the "&" association is shown above the reaction level, indicating that the different proteins form a complex that carries out the reaction. Succinate dehydrogenase is an example of a promiscuous enzyme that can catalyze several different reactions. Isozymes (e.g., GapC and GapA) are

independent proteins that carry out identical reactions. Only one of the isozymes needs to be present for the reaction to occur. Isozymes are shown as two or more arrows leaving different proteins but impinging on the same reaction.

Organism-specific sources of information

Several biological databases that integrate genomic and biochemical data for a particular organism are becoming available. One of the earliest of such sites is the E. coli encyclopedia (EcoCyc) database. Comprehensive Yeast Genome Database (CYGD), Yeast Protein Database (YPD), and Saccharomyces Genome Database (SGD) are some examples for yeast. The widely used Kyoto Encyclopedia of Genes and Genomes (KEGG) database organizes its genomic information as maps of reaction networks.

In reaction maps, arrows are used to connect various metabolites, indicating that one metabolite can be converted to another by a chemical reaction. This representation is the standard graphical representation of reaction and pathway data and will be described in Chapter 6. For many organisms of interest, comprehensive textbooks have been written that include detailed descriptions of the organism's metabolism and other biological functions. These books give an overview of the organism's importance, metabolic features, and important references, as well as physiological data. The *E. coli* two-volume set was the first of its kind and continues to be a useful source when building models of other bacteria. Several such organism-specific compendia have appeared. Such compilations of genetic, biochemical, and physiological data, and functional attributes of a particular organism, represent highly concentrated sources of data needed for reliable reconstructions. In addition, to achieve a high- quality, well-curated network reconstruction, one should search the primary research literature.

Comprehensive review articles are particularly useful since they contain organized collections of primary articles on a particular organism. Reviews are typically well summarized and written by experts on the subject and provide an accessible source of biochemical information. Frequently though, one has to search the primary literature, and searches may have to be done on a regular basis to continually update the network reconstruction.

Models and Modeling

If we observe biological processes, we are confronted with various complex processes that cannot be explained from first principles and the outcome of which cannot reliably be foreseen from intuition. Even if general biochemical principles are well established (e.g., the central dogma of transcription and translation, the biochemistry of enzyme-catalyzed reactions), the biochemistry of individual molecules and systems is often unknown and can vary considerably between species. Experiments lead to biological hypotheses about individual processes, but it often remains unclear if these hypotheses can be combined into a larger coherent picture because it is often difficult to foresee the global behavior of a complex system from knowledge of its parts. Mathematical modeling and computer simulations can help us understand the internal nature and dynamics of these processes and to arrive at predictions about their future development and the effect of interactions with the environment.

What is a Model?

The answer to this question will differ among communities of researchers. In a broad sense, a model is an abstract representation of objects or processes that explains features of these objects or processes (Figure 1.2). A biochemical reaction network can be represented by a graphical sketch showing dots for metabolites and arrows for reactions; the same network could also be

described by a system of differential equations, which allows simulating and predicting the dynamic behavior of that network. If a model is used for simulations, it needs to be ensured that it faithfully predicts the system s behavior – at least those aspects that are supposed to be covered by the model. Systems biology models are often based on well-established physical laws that justify their general form, for instance, the thermodynamics of chemical reactions; besides this, a computational model needs to make specific statements about a system of interest – which are partially justified by experiments and biochemical knowledge, and partially by mere extrapolation from other systems. Such a model can summarize established knowledge about a system in a coherent mathematical formulation. In experimental biology, the term model is also used to denote a species that is especially suitable for experiments, for example, a genetically modified mouse may serve as a model for human genetic disorders.

Advantages of Computational Modeling

Models gain their reference to reality from comparison with experiments, and their benefits therefore depend on the quality of the experiments used. Nevertheless, modeling combined with experimentation has a lot of advantages compared to purely experimental studies: . Modeling drives conceptual clarification. It requires verbal hypotheses to be made specific and conceptually rigorous.

Modeling highlights gaps in knowledge or understanding. During the process of model formulation, unspecified components or interactions have to be determined. Modeling provides independence of the modeled object.

Time and space may be stretched or compressed ad libitum. Solution algorithms and computer programs can be used independently of the concrete system.

Modeling is cheap compared to experiments.

Models exert by themselves no harm on animals or plants and help to reduce ethical problems in experiments. They do not pollute the environment.

. Modeling can assist experimentation. With an adequate model, one may test different scenarios that are not accessible by experiment. One may follow time courses of compounds that cannot be measured in an experiment. One may impose perturbations that are not feasible in the real system. One may cause precise perturbations without directly changing other system components, which is usually impossible in real systems. Model simulations can be repeated often and for many different conditions.

. Model results can often be presented in precise mathematical terms that allow for generalization. Graphical representation and visualization make it easier to understand the system.

. Finally, modeling allows for making well-founded and testable predictions. The attempt to formulate current knowledge and open problems in mathematical terms often uncovers a lack of knowledge and requirements for clarification.

Furthermore, computational models can be used to test whether proposed explanations of biological phenomena are feasible. Computational models serve as repositories of current knowledge, both established and hypothetical, about how systems might operate. At the same time, they provide researchers with quantitative descriptions of this knowledge and allow them to simulate the biological process, which serves as a rigorous consistency test.

UNIT-III

SYLLABUS

Unit – III

Systems Biology Databases KEGG (Kyoto Encyclopedia of Genes and Genomes). BRENDA (BRaunschweig ENzyme DAtabase). BioSilico. EMP (Embden-Meyerh of Parnas). MetaCyc and AraCyc. SABIO-RK (System for the Analysis of Biochemical Pathways - Reaction Kinetics). BioModels.

KEGG (Kyoto Encyclopedia of Genes and Genomes)

is a knowledge base for systematic analysis of gene functions, linking genomic information with higher order functional information. The genomic information is stored in the GENES database, which is a collection of gene catalogs for all the completely sequenced genomes and some partial genomes with up-to-date annotation of gene functions. The higher order functional information is stored in the PATHWAY database, which contains graphical representations of cellular processes, such as metabolism, membrane transport, signal transduction and cell cycle. The PATHWAY database is supplemented by a set of ortholog group tables for the information about conserved subpathways (pathway motifs), which are often encoded by positionally coupled genes on the chromosome and which are especially useful in predicting gene functions. A third database in KEGG is LIGAND for the information about chemical compounds, enzyme molecules and enzymatic reactions. KEGG provides Java graphics tools for browsing genome maps, comparing two genome maps and manipulating expression maps, as well as computational tools for sequence comparison, graph comparison and path computation. The KEGG databases are daily updated and made freely available (http://www.genome.ad.jp/kegg/).

While the genome sequencing projects rapidly determine gene catalogs for an increasing number of organisms, functional annotation of individual genes is still largely incomplete. KEGG (Kyoto Encyclopedia of Genes and Genomes) is an effort to link genomic information with higher order functional information by computerizing current knowledge on cellular processes and by standardizing gene annotations. Generally speaking, the biological function of the living cell is a result of many interacting molecules; it cannot be attributed to just a single gene or a single molecule (1). The functional assignment in KEGG is a process of linking a set of genes in the genome with a network of interacting molecules in the cell, such as a pathway or a complex, representing a higher order biological function.

The KEGG project was initiated in May 1995 under the Human Genome Program of the Ministry of Education, Science, Sports and Culture in Japan (2). All the data in KEGG and associated software tools are made available as part of the Japanese GenomeNet service (3). KEGG consists of three databases: PATHWAY for representation of higher order functions in terms of the network of interacting molecules, GENES for the collection of gene catalogs for all the completely sequenced genomes and some partial genomes, and LIGAND (4) for the collection of chemical compounds in the cell, enzyme molecules and enzymatic reactions. The overall architecture of the KEGG system is basically the same as previously reported (5). The user may enter the KEGG system top-down starting from the pathway (functional) information or bottom-up starting from the genomic information at the KEGG table of contents page (http://www.genome.ad.jp/kegg/kegg2.html).

BRENDA (BRaunschweig ENzyme DAtabase

represents a comprehensive collection of enzyme and metabolic information, based on primary literature. The database contains data from at least 83 000 different enzymes from 9800 different organisms, classified in ~4200 EC numbers. BRENDA includes biochemical and molecular information on classification and nomenclature, reaction and specificity, functional parameters, occurrence, enzyme structure, application, engineering, stability, disease, isolation and preparation, links and literature references. The data are extracted and evaluated from ~46 000 references, which are linked to PubMed as long as the reference is cited in PubMed. In the past year BRENDA has undergone major changes including a large increase in updating speed with >50% of all data updated in 2002 or in the first half of 2003, the development of a new EC-tree browser, a taxonomy-tree browser, a chemical substructure search engine for ligand structure, the development of controlled vocabulary, an ontology for some information fields and a thesaurus for ligand names. The database is accessible free of charge to the academic community at http://www.brenda.uni-koeln.de.

The development of BRENDA was begun in 1987 at the German National Research Center for Biotechnology (GBF) and continues at the Cologne University Bioinformatics Centre. Initially, BRENDA was published as a series of books (1). Since 1998 all data have been presented in a relational database system with access free to the academic community at http://www.brenda.uni-koeln.de. Commercial users are required to purchase a licence.

Enzymes represent the largest and most diverse group of all proteins, catalysing all chemical reactions in the metabolism of all organisms. They play a key role in the regulation of metabolic steps within the cell. With the development and progress of projects of structural and functional

genomics and metabolomics, the systematic collection, accessibility and processing of enzyme data becomes even more important in order to analyse and understand biological processes.

BRENDA, a protein function database (2) contains a huge amount of enzymic and metabolic data and is updated and evaluated by extracting information from the primary literature. Since 2002 the annotation speed has been tripled to 1000 EC numbers per year.

Major developments in the past few years were the ongoing conversion from an EC number/organism-specific to a protein-molecule-specific database. Furthermore, the presentation and the advanced search engine via the World Wide Web was improved. Tools like an EC browser, the taxonomy browser and a sequence-based search engine were integrated. BRENDA now provides the opportunity to search for substructures of ligands and a thesaurus of those chemical compounds that are involved in enzyme reactions. In terms of systematic access and analysis of data, a controlled vocabulary for organism-specific information, i.e. intracellular localization and enzyme source, was established.

BioSilico

is a web-based database system that facilitates the search and analysis of metabolic pathways. Heterogeneous metabolic databases including LIGAND, ENZYME, EcoCyc and MetaCyc are integrated in a systematic way, thereby allowing users to efficiently retrieve the relevant information on enzymes, biochemical compounds and reactions. In addition, it provides well-designed view pages for more detailed summary information. BioSilico is developed as an extensible system with a robust systematic architecture.

Glycolysis or EMP (Embden-Meyerhof-Parnas) pathway

All the living organisms whether aerobes or anaerobes, initiate the mechanism of respiration by breaking down glucose (6 carbon compound) into two molecules of pyruvate (3 carbon compound).

This initial process that occurs in 10 steps, the first 5 of which constitute the preparatory phase and the last 5 constitute the payoff phase is called glycolysis.

The process of glycolysis was first described by Gustav Embden, Otto Meyerhof and Parnas and therefore also referred to as EMP pathway.

The Preparatory phase:

Phosphorylation of glucose:

It is the first priming reaction in glycolysis.

Glucose molecule is phosphorylated in the presence of ATP to form glucose-6-phosphate. This reaction is catalyzed by the enzyme hexokinase which requires a divalent Mg++ as cofactor. ATP (phosphoryl donor) is converted to ADP in the reaction.

Isomerization of glucose-6-phosphate to fructose-6-phosphate:

The enzyme phosphohexose isomerase (phosphoglucose isomerase) catalyzes the reversible isomerization of glucose-6-phosphate, an aldose, to fructose-6-phosphate, a ketose.

Fructose-6-phosphate may be formed directly from free fructose by its phosphorylation in the presence of an enzyme fructokinase, Mg++ and ATP.

Phosphorylation of fructose-6-phosphate to fructose1, 6-biphosphate:

It is the second priming reaction in glycolysis.

Fructose-6-phosphate combines with another phosphoryl group from another ATP molecule, yielding fructose1,6-biphosphate catalyzed by phosphofructokinase-1 (PFK-1) in the presence of Mg++.

Cleavage of frucose1,6-biphosphate:

The enzyme fructose 1, 6-biphosphate aldolase, often called simply aldolase, catalyzes a reversible aldol condensation.

Fructose 1,6- biphosphate is cleaved to yield two different 3-carbon (triose) phosphates, glyceraldehyde 3-phosphate (GAP), an aldose and dihydroxyacetone phosphate (DHAP), a ketose.

Interconversion of two triose phosphates:

Only one of the two triose phosphates formed by aldolase, can be directly degraded in the subsequent steps of glycolysis.

DHAP is rapidly and reversibly converted into GAP with the aid of enzyme phosphotriose isomerase or triose phosphate isomerase.

The Payoff phase:

6. Oxidation of glyceraldehyde3-phosphate to 1,3-biphosphoglycerate:

Glyceraldehyde 3-phosphate combines with a phosphate group (derived from H3PO4 in the cytosol and not from ATP)) and is oxidized with the separation of two atoms of hydrogen from it to form 1,3-biphosphoglycerate which is catalyzed by glyceraldehyde3-phosphate dehydrogenase.

Of the two hydrogen separated, one complete hydrogen atom (proton and electron) and one additional electron are picked up by NAD+ which gets reduced to NADH. The remaining one hydrogen proton or ion (H+) remains free in the cytosol.

Phosphoryl transfer from 1,3-biphospohglycerate to ADP:

High energy phosphate group on carbon 1 of 1,3-biphosphoglycerate is transferred to a molecule of ADP, converting it into an ATP molecule.

1,3-biphosphoglycerate changes to 3-phosphoglycerate due to loss of a phosphate group.

This reaction is catalyzed by the enzyme phosphoglycerate kinase in the presence of Mg++.

Formation of ATP directly from metabolites (substrate) is known as substrate level phosphorylation.

Conversion of 3-phosphoglycerate to 2-phosphoglycerate (isomerization):

Phosphate group on the third carbon of 3-phosphoglycerate shifts to the second carbon, producing 2-phosphoglycerate.

This change is aided by the enzyme phosphoglycerate mutase in the presence of Mg++.

Dehydration of 2-phosphoglycerate to phosphoenol pyruvate:

2-phosphoglycerate loses a water molecule (reversibly)in the presence of an enzyme, enolase and Mg++, and changes into phosphoenol pyruvate (PEP).

Transfer of the phosphoryl group from phosphoenol pyruvate to ADP:

The last step in glycolysis is the transfer of the phosphoryl group from phosphoenol pyruvate to ADP, catalyzed by pyruvate kinase.

This process requires K+ and either Mg++ or Mn++.

AraCyc

is a tool for visualizing biochemical pathways of Arabidopsis thaliana. It is supported by the Pathway Tools software developed by Peter Karp's group at SRI.

AraCyc was originally computationally predicted for the sequenced Arabidopsis genome using MetaCyc as reference database. The existence of the predicted pathways was then manually validated and non-Arabidopsis pathways were removed. The manual curation of the database, which includes correcting pathways and adding missing pathways, is on-going.

AraCyc contains a mix of information: extracted from peer-reviewed literature and computationally predicted. A series of icons usually located in the top-right corner of the screen will indicate the type of evidence used for the displayed pathway or enzyme information displayed (for more info, visit the AraCyc Tutorial).

AraCyc is released on a semi-annual basis (see Release Notes). With each release a Summary of the DataBase Content is made available, as well as a PathoLogic Software Report.

In its current state many AraCyc pathways contain pathways with incomplete reactions and enzyme/gene assignments as well as missing some pathways. We, therefore, strongly encourage AraCyc users to share their pathway knowledge with us to help us in our constant task of improving its data content (see the Data Submission page).

MetaCyc Metabolic Pathway Database

MetaCyc is a curated database of experimentally elucidated metabolic pathways from all domains of life. MetaCyc contains 2642 pathways from 2941 different organisms.

MetaCyc contains pathways involved in both primary and secondary metabolism, as well as associated metabolites, reactions, enzymes, and genes. The goal of MetaCyc is to catalog the universe of metabolism by storing a representative sample of each experimentally elucidated pathway.

MetaCyc applications include:

Online encyclopedia of metabolism

Predict metabolic pathways in sequenced genomes

Support metabolic engineering via enzyme database

Metabolite database aids metabolomics research

SABIO-RK (http://sabiork.h-its.org/)

is a manually curated database containing data about biochemical reactions and their reaction kinetics. The data are primarily extracted from scientific literature and stored in a relational database. The content comprises both naturally occurring and alternatively measured

biochemical reactions and is not restricted to any organism class. The data are made available to the public by a web-based search interface and by web services for programmatic access. In this update we describe major improvements and extensions of SABIO-RK since our last publication in the database issue of Nucleic Acid Research (2012). (i) The website has been completely revised and (ii) allows now also free text search for kinetics data. (iii) Additional interlinkages with other databases in our field have been established; this enables users to gain directly comprehensive knowledge about the properties of enzymes and kinetics beyond SABIO-RK. (iv) Vice versa, direct access to SABIO-RK data has been implemented in several systems biology tools and workflows. (v) On request of our experimental users, the data can be exported now additionally in spreadsheet formats. (vi) The newly established SABIO-RK Curation Service allows responding to specific data requirements.

In 2006, SABIO-RK database (1) has been established to support modelers of biochemical reactions and complex networks. SABIO-RK represents a repository for structured, curated and annotated data about reactions and their kinetics. The data are manually extracted from the scientific literature and stored in a relational database. As compared with automatic data extraction by text mining tools, the manual extraction process guarantees a very high degree of accurateness and completeness. Especially, the extraction of the complex information of reactions and kinetics of most of the available publications are not enough structured and well written. Furthermore, relevant information is distributed over the entire article and unique identifiers or controlled vocabularies are missing (2,3). Based on the time consuming process of manual data extraction and manual curation, SABIO-RK emphasizes on quality rather than on quantity. SABIO-RK is not only a database for modelers but also for experimentalists in the laboratory who are looking for example for more details about the enzymatic activity of a protein

or about alternative reactions of an enzyme. For many years SABIO-RK was focussing on kinetics of metabolic reactions but with an increased user interest in kinetics data for signalling events, SABIO-RK also stores reactions and binding events of signal transduction pathways.

The bidirectional cross-references between SABIO-RK and protein specific databases like UniProtKB, pathway databases like KEGG or chemical compound databases like ChEBI assist users to find more specific kinetic information in SABIO-RK and vice versa.

A comparable database providing kinetic parameters is BRENDA. In contrast to SABIO-RK, the information in BRENDA is centred on enzymes and their kinetic constants, whereas SABIO- RK focuses on reactions and additionally, beside constants, offers the associated kinetic rate laws, formulas and experimental conditions. Other databases containing kinetic data are focussing e.g. on proteins (UniProtKB), plant metabolism (MetaCrop or protein interactions (KDBI (9)).

BioModels

The beta version of BioModels (currently hosted at https://wwwdev.ebi.ac.uk/biomodels/) will be available from https://www.ebi.ac.uk/biomodels. Although some internal links used to load static content such as images or style sheets in classic BioModels (https://www.ebi.ac.uk/biomodels-main) will move, we expect this to be transparent to users. Access to the models, the Model of the Month entries or our documentation pages will be unaffected. These changes will come into effect Monday, 25 June at 1pm GMT.

A redirection from https://wwwdev.ebi.ac.uk/biomodels/ to https://www.ebi.ac.uk/biomodels/ will be in operation for 30 days to allow users to update their bookmarks. Models should be

submitted to https://www.ebi.ac.uk/biomodels. The development instance of BioModels will continue to see new features and general improvements before being rolled out to the main site.

The BioModels US mirror hosted by the California Institute of Technology (http://biomodels.caltech.edu) will be updated on Monday, 2 July between 8am and 10am GMT. During this maintenance period the mirror will be offline and all requests will be redirected to the EBI instance of BioModels.

UNIT-IV

SYLLABUS

Unit – IV

Tools for System Biology: Cell Designer. Ali Baba. Cell Profiler. JDesigner. Bio-SPICE (Biological Simulation Program for Intra and Inter Cellular Evaluation). SBML (Systems Biology Markup Language). SBGN (Systems Biology Graphical Notation). SBML-SAT (SBML based Sensitivity Analysis Tool).

Cell Designer

is a structured diagram editor for drawing gene-regulatory and biochemical networks. Networks are drawn based on the process diagram, with graphical notation system proposed by Kitano, and are stored using the Systems Biology Markup Language (SBML), a standard for representing models of biochemical and gene-regulatory networks. Networks are able to link with simulation and other analysis packages through Systems Biology Workbench (SBW). Cell Designer supports simulation and parameter scan by an integration with SBML ODE Solver, SBML Simulation Core and Copasi.

By using Cell Designer, you can browse and modify existing SBML models with references to existing databases, simulate and view the dynamics through an intuitive graphical interface.

ALI BABA: A TEXT MINING TOOL FOR SYSTEMS BIOLOGY

Text mining is the process of automatically deriving information from text (as opposed to data mining that works on structured data). This process starts with accessing the relevant literature and ends with extracting the desired pieces of information. Access mostly is provided by Web-based search tools, the best known of which is PubMed. PubMed currently contains citations from close to 18 million publications in the biomedical domain (biology, biochemistry, medicine, and related fields), from approximately 5200 journals, since 1865. Up to 4000 citations (abstract and bibliographical information) are added to PubMed per day, which necessitates automated means to efficiently handle searches for high-quality information

The Cell Profiler 2.1.0 release enhances Cell Profiler in eight general areas:

New input modules and project storage files. Cell Profiler has four new modules—Images, Metadata, Names And Types and Groups—that are designed to assemble Cell Profiler's image sets in a more flexible and intuitive form. These four modules operate on a list of paths to image files, extracting metadata, filtering and grouping them into channels. The file list and pipeline are now stored in a project file. Cell Profiler 2.1.0 assembles large image sets much more quickly than its predecessor and caches image sets in many circumstances to allow quick start-up times for pipelines. The legacy input modules, Load Images and Load Single Images can still be used as before and the Load Data module can both be used as before and can be used with an image set list exported from a project.

Multi-core processing. Cell Profiler now utilizes multiple cores in analysis mode, which allows for faster processing without any user intervention necessary. It starts a number of worker processes and partitions its work among them. In addition to the increased utilization of CPU resources, this multiprocessing mode allows the workers to run in a 64-bit address space on

OS/X even though the UI process is constrained to a 32-bit address space. Cell Profiler is still optimized to run on a single core when headless.

Measurements saved to disk as HDF5 files during processing. Measurements are now preferentially stored as HDF5 files which are written to disk during the course of the analysis. Previously, measurements were stored in memory and written to a Matlab format file at the end of the run. The change lowers Cell Profiler's memory footprint, especially for long runs, provides incremental measurement output and has paved the way for alternative uses of the HDF5 file.

Improved support for Linux. We provide Linux release and trunk builds for CentOS and Ubuntu.

Improved help. The help and welcome screen have been extensively revised to make them easier to use and more complete. The help is now searchable.

GIT and Github. In the interim between the Cell Profiler 2.0 (11710) and the 2.1.0 release, we've moved our version control from SVN to GIT (https://github.com/CellProfiler/CellProfiler) and we now use the Github repository issue tracking system to file and resolve issues. We welcome Github pull requests.

Other added functionality. We have added new image operations, thresholding methods, and measurements. Cell Profiler can load images from OMERO. Users can take advantage of the extended functionality of Image J 2.0 in addition to Image J 1.0 support.

JDesigner

is a graphical modeling environment for biochemical reaction networks. It allows drawing the network on screen and selecting the appropriate kinetic laws from a wide selection of inbuilt rate

laws or defining new user defined rate laws. JDesigner is probably the module that profits the most from SBW. Via the previously described auto-layout and SBML Layout extension modules, it is able to render any SBML file. Using the SBW menu, it allows passing on the current model to a wide variety of analysis tools. It also allows for time-course simulation and steady state analysis using either roadRunner or Jarnac. Furthermore, it uses Metatool for structural analysis. Finally, it dynamically finds modules from the SBML Exporter category at runtime to allow exporting the model into a variety of file formats including Matlab, Java, or XPP.

Bio-SPICE

Biological Simulation Program for Intra- and Inter-Cellular Evaluation

Bio-SPICE, an open source framework and software toolset for Systems Biology, is intended to assist biological researchers in the modeling and simulation of spatio-temporal processes in living cells. In addition, our goal is to develop and serve a user community committed to using, extending, and exploiting these tools to further our knowledge of biological processes.

In collaboration with other Bio-SPICE Community members, we will develop, license, distribute, and maintain a comprehensive software environment that integrates a suite of analytical, simulation, and visualization tools and services to aid biological researchers engaged in building computable descriptions of cellular functions. From disparate data analysis and information mining to experimental validation of computational models of cell systems, our environment will offer a comprehensive substrate for efficient research, collaboration, and publication.

Mission

Bio-SPICE is intended for modeling and simulation of spatio-temporal processes in living cells.

The goals of Bio-SPICE are to support discovery through:

Developing computational and mathematical models of bio-molecular systems in cells capturing the nature of gene-protein interactions

Developing tools that can rapidly incorporate relevant experimental data and knowledge known in the literature to build models of pathways, networks, and spatial processes

Developing simulation tools for the dynamic analysis of bio-molecular systems

Creating an extensible framework for easy insertion of models and their refinement, as well as customization to specific mechanisms

In addition, our goal is to develop and serve a user community committed to using, extending, and exploiting these tools to further our knowledge of biological processes. In collaboration with other Bio-SPICE Community members, we will develop, license, distribute, and maintain a comprehensive software environment that integrates a suite of analytical, simulation, and visualization tools and services to aid biological researchers engaged in building computable descriptions of cellular functions. From disparate data analysis and information mining to experimental validation of computational models of cell systems, our environment will offer a comprehensive substrate for efficient research, collaboration, and publication.

Background

The biological sciences have experienced dramatic growth over the past two decades, culminating in the recent completion of a working draft of the human genome. This stands as a seminal scientific achievement and is indicative of the growing repertoire of available molecular,

biochemical, and physiological approaches by which investigators can elucidate the complex inner-workings of living cells. Research hypotheses that were unstable only a short time-period ago are fast becoming amenable to the scientific method.

Originally a DARPA-sponsored program, the goal of Bio-SPICE was to apply expertise from the computational and information sciences in order to develop in silico simulations that model how various pathogens act to disrupt normal cellular processes. The program was transitioned to an Open Source project at the conclusion of its DARPA sponsorship in December 2006. The project is now hosted by SourceForge.net. Originally developed under the DARPA BioCOMP license, Bio-SPICE is now released under the open-source BSD license. (See license below.)

Original key contributors to Bio-SPICE included: Virginia Polytechnic University, University of Pennsylvania, University of Texas, Harvard University, University of California, Los Angeles and Berkeley, Lawrence Berkeley National Laboratory, California Institute of Technology, New York University, Massachusetts Institute of Technology, and the Molecular Sciences Institute. The lead integrator for the project was SRI International.

Original DARPA Bio-SPICE release announcement

Bio-SPICE BSD License

Selected publications of the Bio-SPICE project

OMICS issue 1

OMICS issue 2

The Systems Biology Markup Language (SBML)

is a representation format, based on XML, for communicating and storing computational models of biological processes. It is a free and open standard with widespread software support and a community of users and developers. SBML can represent many different classes of biological phenomena, including metabolic networks, cell signaling pathways, regulatory networks, infectious diseases, and many others. It has been proposed as a standard for representing computational models in systems biology today.

Motivation: Molecular biotechnology now makes it possible to build elaborate systems models, but the systems biology community needs information standards if models are to be shared, evaluated and developed cooperatively.

Results: We summarize the Systems Biology Markup Language (SBML) Level 1, a free, open, XML-based format for representing biochemical reaction networks. SBML is a software-independent language for describing models common to research in many areas of computational biology, including cell signaling pathways, metabolic pathways, gene regulation, and others.

Availability: The specification of SBML Level 1 is freely available from http://www.sbml.org/

The Systems Biology Graphical Notation.

Circuit diagrams and Unified Modeling Language diagrams are just two examples of standard visual languages that help accelerate work by promoting regularity, removing ambiguity and enabling software tool support for communication of complex information. Ironically, despite having one of the highest ratios of graphical to textual information, biology still lacks standard graphical notations. The recent deluge of biological knowledge makes addressing this deficit a pressing concern. Toward this goal, we present the Systems Biology Graphical Notation (SBGN),

a visual language developed by a community of biochemists, modelers and computer scientists. SBGN consists of three complementary languages: process diagram, entity relationship diagram and activity flow diagram. Together they enable scientists to represent networks of biochemical interactions in a standard, unambiguous way. We believe that SBGN will foster efficient and accurate representation, visualization, storage, exchange and reuse of information on all kinds of biological knowledge, from gene regulation, to metabolism, to cellular signaling.

SBML-SAT

This tool is designed to implement a variety of simulation, sensitivity analysis, steady state analysis and robustness analysis for ordinary differential equations (ODE) based biological models including biophysical models, signaling pathways, gene regulation networks and metabolic pathways.

SBML-SAT supports the import of models in the System Biology Mark- up Language (SBML) format.

The following is a summary of the features/capabilities of SBML-SAT:

Support for SBML import and export --

Current release of SBML-SAT supports:

(1) SBML Level 1 Version 1 and 2.

(2) SBML Level 2 Versions 1, 2 and 3.

The import of a model to SBML-SAT should be an SBML Extensible Markup Language (XML) file. SBML-SAT will automatically generate the necessary ODE models and run the simulation and various analyses.

Support of events and rules -- SBML-SAT supports rules (except algebraic rules) and events very well including different event(s) triggers (complicated logical triggers are supported) and assignments.

Local sensitivity analysis -- SBML-SAT provides traditional local sensitivity analysis for the SBML models.

Local sensitivity analysis is the study of the changes in the model outputs with respect to parameter (factor) variations around a local point in the parameter space, which are quantified by the sensitivity coefficients.

Mathematically, the sensitivity coefficients are the first order derivatives of model outputs with respect to the model parameters.

Global sensitivity analysis -- SBML-SAT can run four (4) different global sensitivity analysis algorithms for the SBML-Models, which are:

1) Multi-Parametric Sensitivity Analysis (MPSA) - MPSA can be used to study the relative importance of the parameters with respect to the model output.

The basic idea of MPSA is to map the uncertainty of the parameters into the model output by randomly generating parameter values from predefined distributions (without prior knowledge, uniform distributions are assumed).

2) Partial Rank Correlation Coefficient Analysis (PRCC) - The PRCC method is a rank transformed linear regression analysis that is routinely used for analysis of systems with a nonlinear and monotonic relationship between the system inputs and outputs. Linear regression analysis methods best fit a straight line to input and output values.

When nonlinear, monotonic relationships exist between system input and outputs, poor linear regression fits can be alleviated by performing the linear regression analysis on a rank ordered list of the model output and input values.

PRCC calculates the sensitivity indices from the Pearson correlation coefficients between the model output and input parameters as well as each pair of parameters after rank transformation.

(3) SOBOL's Method - is a variance based method that makes No assumptions on the relationship between the system inputs and outputs. It is computationally expensive since it utilizes a large number of model simulations with parameter values sampled from the parameter space by the winding stair algorithm.

The variance of the numerous model outputs is estimated by Monte Carlo integrations. The model output variance is apportioned into summands of partial variances from combinations of input parameters with increasing dimensionality.

The total effects sensitivity indices quantify all of the effects that a parameter, in combination with any other parameter(s), has on the model output.

They are defined as the ratio of the sum of the related partial variances to the overall variance of the model output. The larger the fraction, the higher is the corresponding sensitivity. SBML-SAT calculates the 'total effect sensitivity' indices.

(4) Weighted Average of Local Sensitivities - In this approach, local sensitivity indices are calculated at multiple random points within the parameter space; a weighted average of the local sensitivity indices serves to provide some approximation of the global parameter sensitivities.

Steady state analysis -- SBML-SAT runs steady state analysis for the stable system. The steady state of the model can be detected only when all the components (state variables) of the model don't change at a certain time.

Robustness analysis -- Robustness is one of the fundamental properties of biological systems, which allows the system to maintain its behavior against random perturbations.

SBML-SAT implements robustness analysis for a variety of model output by simultaneous variations of the selected parameters/initial conditions.

SBML model editor module -- Currently, a SBML model editor module is Not available in SBML-SAT. Fortunately, many existing free software packages such as CellDesigner (see G6G Abstract Number 20159), SBML editor and COPASI, share a common functionality for constructing and editing SBML models.

The users can generate their models with these free software packages and then run a variety of analyses in SBML-SAT by importing the model in SBML format.

Although SBML-SAT doesn't provide a SBML editor for model construction, it provides a convenient track for modifying the initial conditions of the state variables and parameter values in the model.

UNIT-V

SYLLABUS

Unit – V

Premises & Promises of Systems Biology: Premise of Systems Biology. Promise of Systems Biology. Challenges of Systems Biology. Applications of Systems Biology.

Systems Biology and Cancer: Promises and Perils

Systems biology uses systems of mathematical rules and formulas to study complex biological phenomena. In cancer research there are three distinct threads in systems biology research: modeling biology or biophysics with the goal of establishing plausibility or obtaining insights, modeling based on statistics, bioinformatics, and reverse engineering with the goal of better characterizing the system, and modeling with the goal of clinical predictions. Using illustrative examples we discuss these threads in the context of cancer research.

The term "systems biology" has different meanings for different groups of researchers. One view is that "systems biology is a powerful new paradigm based on the premise that the properties of complex systems consisting of many components that interact with each other in non-linear, non-additive ways cannot be understood solely by focusing on the components'. A reviewer of this manuscript defined systems biology as "an interdisciplinary study of biological systems that spans multiple scales in space and time." In our view, systems biology is the study of biological phenomena using systems of mathematical rules.

or formulas.

There is a large literature involving systems biology and actual or potential applications to cancer research (e.g., the review article by Laubenbacher et. al., 2009, and other references in that paper and here). What is sometimes missing is a perspective that sees various threads of systems biology as different parts of a single fabric. In particular, we see a divide between systems biology based on the principles of biology or biophysics, systems biology related to statistics, bioinformatics, and reverse engineering, and systems biology involving clinical predictions, sometimes without full appreciation of other viewpoints. To provide perspective, we briefly discuss these three major threads in system biology and their relation to cancer research. This is not a comprehensive review, but rather draws upon selected articles as illustrative examples.

Computational challenges in systems biology

Systems biology is a broad field that incorporates both computational and experimental approaches to provide a system level understanding of biological function. Initial forays into computational systems biology have focused on a variety of biological networks such as protein–protein interaction, signaling, transcription and metabolic networks. In this review we will provide an overview of available data relevant to systems biology, properties of biological networks, algorithms to compare and align networks and simulation and modeling techniques. Looking towards the future, we will discuss work on integrating additional functional information with biological networks, such as three dimensional structures and the complex environment of the cell. Combining and understanding this information requires development of

novel algorithms and data integration techniques and solving these difficult computational problems will advance both computational and biological research.

Systems Medicine: The Application of Systems Biology Approaches for Modern Medical Research and Drug Development

The exponential development of highly advanced scientific and medical research technologies throughout the past 30 years has arrived to the point where the high number of characterized molecular agents related to pathogenesis cannot be readily integrated or processed by conventional analytical approaches. Indeed, the realization that several moieties are signatures of disease has partly led to the increment of complex diseases being characterized. Scientists and clinicians can now investigate and analyse any individual dysregulations occurring within the genomic, transcriptomic, miRnomic, proteomic, and metabolomic levels thanks to currently available advanced technologies. However, there are drawbacks within this scientific brave new age in that only isolated molecular levels are individually investigated for their influence in affecting any particular health condition. Since their conception in 1992, systems biology/medicine focuses mainly on the perturbations of overall pathway kinetics for the consequent onset and/or deterioration of the investigated condition/s. Systems medicine approaches can therefore be employed for shedding light in multiple research scenarios, ultimately leading to the practical result of uncovering novel dynamic interaction networks that are critical for influencing the course of medical conditions. Consequently, systems medicine also serves to identify clinically important molecular targets for diagnostic and therapeutic measures against such a condition.

The exponential development of highly advanced scientific and medical research analytical technologies throughout the past 30 years has arrived to the point where most (if not all) key molecular determinants deemed to affect human conditions and diseases can be scrutinized with great detail.

Scientists and clinicians can now begin to attempt investigation of any individual dysregulations occurring within the genomic, transcriptomic, miRnomic, proteomic, and metabolomic levels thanks to advancing wet-lab technologies such as mass spectrometry, quantitative polymerase chain reaction (qPCR) and next generation sequencing, and detailed bioinformatics suites. All these technologies are capable of extracting information from complex datasets to enable disease models to be developed for wet-lab testing. The interplay between the wet and dry lab with specific clinical expertise not only is a main current component of translational medicine, but also is enabled by systems medicine.

However, there are drawbacks within this scientific brave new age, in that in most scientific studies it is only specific molecular levels which are individually investigated for their influence in affecting any particular health condition. Ideally, any form of medical research with the scope of rooting out dysregulated molecular pathway interactions should focus on investigating the holistic aspects of the complex and multifactorial medical condition/s. This involves careful and methodical examination of all simultaneous molecular interactions occurring levels (e.g., genomic, transcriptomic, etc.). Such "bigger-picture" research perspectives lead to a higher level of understanding for complex and multifactorial disease conditions and ultimately "fast-track" the identification and clinical diagnosis of specific molecular pathway dysregulations with

pathogenesis value, together with the combined identification of novel drug targets for the development of effective translational therapeutics for the medical condition.

Consequently, the urgent need to counteract such research shortcomings has been acquiesced through the emergence, in the last decade, of the novel research field of systems biology

Application of Systems Biology for Human Disease: The Advent of Systems Medicine

The traditional reductionist approach to medical research has been discussed and can be restricted to the investigation of the biological effects of individual or minute quantities of key molecular players for complex, multifactorial human conditions, including cancer. The application of systems biology within the remit of present day medical research can be defined as systems medicine, its concept dating back to 1992 [9]. Such a wider perspective opens new doors of perception and insight into the holistic nature of such disease conditions, focusing mainly on the perturbations of overall pathway kinetics for the consequent onset and/or deterioration of the investigated condition/s. Systems medicine requires the employment of several vital facets in order to attain its clinical theranostic goals whenever such an approach is implemented [35] (see Figure 2).

The essential facets of systems medicine should ideally be established in order to provide proper support for the effective and rapid implementation of any novel research methodologies aimed at reaching the intended outcome for systems medicine-based projects. Undoubtedly, the laboratories involved in conducting systems medicine projects should have the necessary infrastructure and research protocol adaptations required for the intense interdisciplinary networking and consequent data handling and flow of information that are vital components for enabling successful systems medicine approaches.

Another important component for systems medicine involves the employment of computation of computational and modeling sciences. Such expertise is a prerequisite for the effective handling of "big" datasets and also for the interpretation of wet laboratory data in terms of the development of complex interrelationships between varying key molecular players.

Neuroblastoma is the first human condition that has been investigated from a systems level perspective in recent years [19]. Logan et al. constructed a regulatory network model for the main oncogene in neuroblastoma, MYCN, and consequently evaluated the perturbation of this model through the introduction of retinoid drugs (fenretinide, 13-cis-retinoic acid), therefore allowing enhanced insight into the responses of NB tumours to retinoid therapy through the identification of novel molecular interaction hypotheses that can be put to the test in a laboratory setting [19].

The study conducted by Sarmady et al. is apt in demonstrating the versatility of systems-based computational analysis on previously existing experimental data from specific molecular interactions, for the purpose of identifying key molecular players affecting the pathogenesis and severity of the disease condition, in this case Human Immunodeficiency Virus (HIV) [36]. The study applied a motif discovery algorithm on specific groups of HIV viral protein sequences, together with the sequences of immediate binding protein partners found on the host organism [36]. This algorithm ultimately selected only those statistically enriched motifs with conserved viral sequences binding to targeted host proteins [36]. Such an interactome and sequence-based prediction methods allowed for the elucidation of the HIV Nef protein as the main minding site to a multitude of host proteins such as MAPK1, VAV1, LCK, HCK, HLA-A, CD4, FYN, and GNB2L1 [36].

Another example for the use of modeling sciences in systems medicine would be the study conducted by Verma et al., which constructed a systems-based protein regulatory network for the effects of microRNAs (miRNAs) influencing BCR.ABL oncoprotein expression and phosphorylation levels within chronic myeloid leukaemia imatinib-resistant cell line models [37]. This protein regulatory network was deemed to be reliable to identify the varying effects of two specific classes of drugs (tyrosine kinase inhibitors and BCR.ABL-specific miRNAs) on cell lines with differing expression profiles and chemoresistance properties [37]. In addition, for the purpose of this study, quantitative PCR-based high-throughput miRNA expression profiles were established, exemplifying the use of a systems-based approach to develop a protein regulatory network from large scale experimental datasets [37]. This study can also be utilized to illustrate the importance of quantitative analytical technologies (in this case, high-throughput RT-qPCR) for driving novel data collection within systems medicine research.

An alternative research scenario in which systems medicine approaches are highly valuable is in the field of biomarker discovery [8]. The recent study carried out by Zhang et al. analysed in silico the expression data obtained from high-throughput miRNA and mRNA expression profiling analyses for both primary and metastatic prostate cancer [38]. The results of this analysis highlighted the distinction between two separate miRNA-mRNA correlation and regulatory modular networks for primary and metastatic prostate cancer [38]. This study is a classic example to demonstrate the utility of systems level research for the identification of highly interactive biomarkers delineating differing classes and severity for an individual disease condition that can ultimately serve to monitor (or predict) specific treatment responses in the individual patient.

Another crucial requirement for the successful implementation of systems medicine is the availability of significantly large, though also highly defined, patient groups. Such patient groups can be organized particularly well if patient samples are provided from curated biobanks.

The study conducted by Albrecht et al. investigated the pathogenesis of hyperuricaemia through the analysis of high-throughput metabolomic profiling data for the regulation of serum urate, which directly induces gout condition in humans and is also associated with cardiovascular disease and diabetes type II [39]. This study employed Gaussian Graphical Modeling with a hypothesis-free approach for the analysis of 355 metabolites from a total of 1764 patients, with the intention to construct a metabolite network affecting serum urate production [39]. The results of this study elucidated a novel serum urate regulatory pathway involving 38 key metabolite components, with a high proportion of such components bearing a gender-specific trait [39].

The study carried out by Mani et al. provides further evidence for the valuable role sustained by adopting a systems level approach for the prediction of oncogenes in cancer conditions [40]. This particular study focused on the analysis of the B-Cell interactome and microarray-based datasets to predict novel oncogenes and molecular perturbation targets, through the identification of dysregulated molecular interactions, in three specific non-Hodgkin's lymphomas [40].

The advances in imaging sciences and quantitative data extraction methodologies have also been of immense value in attaining successful outcomes through systems medicine approaches. Examples of such technologies include the advent of high content imaging, laser assisted microdissection, and single cell sequencing technology to name but a few [41–43].

Other medical conditions scrutinized through systems level research methodologies include the proliferative fibromatosis condition known as Dupuytren's Disease [44–46] and breast cancer [47] and also within the remit of immunology research [48].

In essence, this change in research perspective by scrutinizing overall molecular network interactions, rather than individual molecules, allows for more effective and clinically applicable research outcomes.

Systems biology can be defined as	the systematic study of organisms using a set of established guidelines	the detailed study of the components of biochemical pathways, such as receptor proteins, using simulation methods, such as molecular dynamics	the combined study of biological systems integrating experimental and computational methods and focussing on the interactions between components	the thorough study of the dynamics of the glycolysis pathway	the combined study of biological systems integrating experimental and computational methods and focussing on the interactions between components
A model in systems biology is usually not	simplified	a mathematical representation of biological processes	the central element of systems biology research	valid under all conceivable environmental conditions	valid under all conceivable environmental conditions
Ordinary differential equation (ODE) models of biochemical networks	are based on continuously-valued variables for the concentrations of chemical species	describe biochemical systems in terms of probabilities of reactive events	are usually linear	can usually be solved analytically using sophisticated software, such as COPASI	are based on continuously-valued variables for the concentrations of chemical species
Which of the following statements is not correct?	One set of parameters determines unambiguously the variables in the model	One set of variables can potentially be caused by many parameter sets	The values of parameters do not change in the course of a simulation	The fluxes of reactions are usually parameters of the model	The fluxes of reactions are usually parameters of the model
In the iterative modelling cycle the so-called forward	optimisation	parameter fitting/estimation	stochastic simulation	text mining in databases of scientific articles	stochastic simulation

Stoichiometric coefficients indicate	the speed of a reaction	the proportions in which substrates are consumed and products are produced	the change in steady state concentrations upon a small increase of the rate of a reaction	whether a steady state is stable or not	the proportions in which substrates are consumed and products are produced
What is not needed when setting up and simulating an ordinary differential equation	the sensitivities (flux control coefficients)	the structure of the reaction system (stoichiometry)	the species concentrations at time point 0 (inital state)	a description of the velocities of reactions (kinetic functions/rate laws)	the sensitivities (flux control coefficients)
Let us assume the reaction S P is in equilibrium. By how much does the ratio of product concentration to substrate concentration change if	by 2	by 1/2	by 1/3	it does not change at all	it does not change at all
If the system is in a steady state	the system always returns to the steady state after very small perturbations	there can be no other stable steady state with the same parameter set	the system shows periodic changes in species' concentrations	the rate of change of species' concentrations is zero	the rate of change of species' concentrations is zero
A system's steady state is asymptotically stable if	all eigenvalues of the corresponding Jacobian matrix have strictly negative real parts	all eigenvalues of the corresponding Jacobian matrix have strictly positive real parts	there is at least one eigenvalue of the corresponding Jacobian matrix with value greater than 1.0	its stoichiometric matrix does not have full rank	all eigenvalues of the corresponding Jacobian matrix have strictly negative real parts
Which of the following can not be used to find steady states of a system	Newton method	Gauss elimination algorithm	forward integration	backward integration (in reverse time direction)	Gauss elimination algorithm

Conservation relations are	linear combinations of parameters that stay constant during a simulation	nonlinear functions of reaction fluxes that stay constant during a simulation	linear combinations of species' concentrations that stay constant during a simulation	nonlinear functions of species' concentrations that stay constant during a simulation	linear combinations of species' concentrations that stay constant during a simulation
Which of the following is not saved in a file that contains a model in the Systems Biology	initial conditions	kinetic functions	parameters for the simulation method, such as step size	the chemical equations including all stoichiometric coefficients	parameters for the simulation method, such as step size
Annotations in models, as defined in the Minimal Information Requested In the Annotation of Models (MIRIAM) standard,	link (arbitrarily named) model entities to entities in the real world, such as specific proteins	unify the naming of model entities across different models	specify in a unique way how model entities should be displayed in a qualitative model diagram	give ratings for the biological correctness of parts of the model	link (arbitrarily named) model entities to entities in the real world, such as specific proteins
Human Mitochondria was sequenced in the	1982	1987	1981	1986	1981
The first proposal of the human genome	Winfield	David	Churchill	R.Sinsheimer	R.Sinsheimer
NCBI was established	1988	2001	2000	2017	1988
National Institute of health is abbreviated as	NIH	NH	INH	NI	NIH
Craig Venter established	Institute for Genomic Research	Institute for Taxonomic Research	Institute for Mitochindrial Research	Institute for Protein Research	Institute for Genomic Research
Craig Venter invented …….	EST	TSE	ETS	STE	EST
Haemophilius influenzae genome was	1992	1993	1994	1995	1995
E.coli genome was published in the year	1997	1995	1996	1999	1997
E.coli genome contains ……protein	4284	7098	6543	6666	4284

E.coli genome contains ……structural	111	122	133	144	122
The first archeal genome to be	E.coli	Virus	M.jannaschii	Nematodes	M.jannaschii
M.jannaschii contains …… number of	1745	1754	1888	1743	1743
The simplest organism	M.genitalium	C.elegans	HIV virus	Nematodes	M.genitalium
First plant genome	Arabidopsis	Rice	Wheat	Sugarcane	Arabidopsis
The scientific endeavours that aim to determine the complete	Genome project	Organelle project	Proteomics	Metabolomics	Genome project
The approximate number of genes in	10000-15000	15000-20000	20000-25000	25000-50000	25000-50000
Sequence database searching algorithm	Watson	crick	Sanger	Wilbur Lipman	Wilbur Lipman
Expansion for EST	Expressed sequence tag	Expressed short tag	Enriched sequence tag	Enriched short tag	Expressed sequence tag
Rice genome	1990	1995	2000	2005	2005
SSS means	Sequence Search Services	Structure Search Services	Similarity Search Services	Smart Search Services	Sequence Search Services
Working draft of human genome was	1995	2000	2005	2010	2000
Complete draft of human genome was	1995	2000	2003	2005	2003
Human genome project	Francis collins	Ari Patrinos	Sanger	Wilbur Lipman	Ari Patrinos
Human genome project	Francis collins	Ari Patrinos	Sanger	Wilbur Lipman	Francis collins
Number of nucleotides in Human haploid	1 billion	2 billion	3 billion	4 billion	3 billion
EMBL created at	1990	1992	1994	1996	1994
Celera group used the method for human	shot gun	genome walking	cDNA library	ESTs	shot gun
BLAST was introduced	1980	1990	2000	2010	1990
Atlas of Protein sequences developed	Margaret Dayhoff's	Owen White	Francis collins	Ari Patrinos	Margaret Dayhoff's

IHGSC means	International Human Genome sequencing consortium	Indian Human Genome sequencing consortium		Internal Human Genome Sequencing Consortium	International Human Genome sequencing consortium
The percentage of protein coding regions	1.1 to 1.4%	2.0 to 2.5%	3.0 to 4%	5.0 to 6.0%	1.1 to 1.4%
Human genome project was initiated by	NIH and DOE	NIH and DDBJ	NIH	EMBL	NIH and DOE
The largest gene in	Immunoglobin	Dystrophin	Heamoglobin	Antibody	Dystrophin
According to HGP geentic similarity of all	99%	95%	99.99%	98%	99.99%
The first draft of HGP was published in the	Science	Cell	PloS Biology	Nature	Nature
The private company involved in Human genome sequencing in	IBM	TATA	HCL	Celera	Celera
HGP was also focussed on identifying	SNP	VNTR's	Satellite DNA	Junk DNA	SNP
All are genome sequencing strategies except	shot gun	Edman Degradation method	Directed gene sequencing	Whole genome short gun sequenicng	Edman Degradation Method
The term genomics was	Abraham Christophen	Elizabeth Angel	Winston Churchill	Thomas Roder	Thomas Roder
Chromosomal and linkage maps are	family data	RFLP	positional Cloning	Sequenceing	family data
The human genome project began as researchers mapped	RFLP's	lods	PCR	VNTR's	RFLP's
Approximately what proportion of the human genome is made	20%	25%	15%	50%	50%
How many protein coding genes do human	10 - 15,000	30 - 40,000	More than 100,000	20 - 25,000	20 - 25,000
How many chromosomes do	48	44	44	46	46

Genes are made up of	DNA	RNA	Proteins	Enzymes	DNA
The human genome is	Responsible for all our physical characteristics	All of our genes	All of the DNA and RNA in our cells	All of our DNA	All of our DNA
Working draft of human genome was	1999	2000	2001	2004	2000

Which out of the following statements is true about regulation of metabolic pathway?	Most of the metabolic pathways are regulated	Most of the metabolic pathways are not regulated	Regulation of metabolic pathways always involves changing the amount of enzymes	Metabolic regulation always depends on control by hormones	Most of the metabolic pathways are regulated
The rate of breakdown of metabolites is termed as	Metabolic state	Metabolism	Steady state	Homeostasis	Steady state
Diminished delivery of oxygen to tissues is termed as	Hypoxia	Ischemia	Homeostasis	Metabolism	Hypoxia
Diminished flow of blood to tissues is termed as	Hypoxia	Ischemia	Homeostasis	Metabolism	Ischemia
Which of the following statements is true about the control of muscle glycogen phosphorylase?	It is activated by phosphorylation by an active phosphorylase kinase	It is allosterically activated by ATP	It is allosterically activated by cAMP	Normally it exists in active form	It is activated by phosphorylation by an active phosphorylase kinase
Which of the following is not a factor determining the activity of an enzyme?	Association with regulatory protein	Sequestration	Allosteric regulation	Nucleotides	Nucleotides
Which of the following statements is true?	High insulin/glucagon ratio activates lipolysis in muscle	High insulin/glucagon ratio inhibits lipolysis in liver	High insulin/glucagon ratio activates lipolysis in adipocytes	Low insulin/glucagon ratio activates lipolysis in adipocytes	Low insulin/glucagon ratio activates lipolysis in adipocytes
Which of the following type of metabolite is used for generating glucose under severe starvation conditions?	Amino acids	Fats	Glycogen	Starch	Amino acids
Which of the following statements is true about brain metabolism in starvation?	The brain can use glucogenic amino acids for energy	The brain can only use glucose as fuel	Up to a quarter of energy requirement of the brain can come from fatty acids	Up to a half of energy requirement of the brain can come from ketone bodies	The brain can use glucogenic amino acids for energy

One of the following statements about the control of enzyme activity by phosphorylation is correct	Phosphorylation of an enzyme results in conformational change	Phosphorylation of an enzyme occurs only at specific tyrosine residues	Phosphorylation of an enzyme is carried out by phosphoprotein phosphatases	Enzyme control by phosphorylation is irreversible	Phosphorylation of an enzyme results in conformational change
Amino acids are	building block of carbohydrates	buiding block of vitamins	building block of minerals	building block of proteins	building block of proteins
Amino acids have	both amino and carboxyl group	both amino and keto group	amino group only	carbonyl and keto group	both amino and carboxyl group
Which of the following statements is true about a peptide bond	It is non planar	It is capable of forming a hydrogen bond	The cis configuration is favoured over the trans configuration	hydroxy group	It is capable of forming a hydrogen bond
Glycine and proline are the most abundant amino acids in the structure of	Hemoglobin	Myoglobin	Insulin	Collagen	Collagen
Which out of the following amino acids carries a net positive charge at the physiological p H	Valine	Leucine	Isoleucine	amino acids	Leucine
Which out of the following amino acids is a precursor for a mediator of allergies and inflammation	Histidine	Tyrosine	Phenyl Alanine	Tryptophan	Histidine
All of the below mentioned amino acids can participate in hydrogen bonding except one	Serine	Cysteine	Threonine	Valine	Valine
All of the following amino acids are both glucogenic as well as ketogenic except	Isoleucine	Leucine	Tyrosine	Phenyl alanine	Leucine
Which out of the following amino acid is a precursor of niacin (Vitamin)?	Tyrosine	Threonine	Tryptophan	Phenylalanine	Tryptophan
Which of the following peptides is cyclic in nature-?	Glutathione	Gramicidin	Met encephalin	Leuencephalin	Gramicidin
Which out of the followings is not a fibrous protein?	Carbonic anhydrase	Collagen	Fibrinogen	Keratin	Carbonic anhydrase
Which out of the following is not a haemo protein	Catalase	Myeloperoxidase	Glutathione peroxidase	Aconitase	Aconitase

All the below mentioned proteins are metalloproteins except	Carbonic anhydrase	Xanthine oxidase	Lactate dehydrogenase	Superoxide dismutase	Lactate dehydrogenase
Which out of the following is a peptide antibiotic	Erythromycin	Gramicidin	Ciprofloxacin	Tetracycline	Gramicidin
Which of the following amino acids is most compatible with an α- helical structure?	Tryptophan	Alanine	Leucine	Proline	Alanine
The highest concentration of cystine can be found in	Melanin	Keratin	Collagen	Myosin	Keratin
Which of the amino acids below is the uncharged derivative of an acidic amino acid?	Cystine	Tyrosine	Glutamine	Serine	Glutamine
Which of the following amino acids is sweet in taste?	Glycine	Alanine	Valine	Glutamic acid	File transfer protocol
Sulphur containg amino acids are	Cystine and Methionine	Methionine and threonine	Cysteine and threonine	Cysteine and serine	Cystine and Methionine
The 21st amino acid is	hydroxy lysine	hydroxy proline	Selenocysteine	citruline	Selenocysteine
The amino acid used in the SDS PAGE electrophoresis	Aspartic acid	Glutamic acid	Glycine	Aspartic acid and Lysine together	Glycine
Single letter code of pyrrolysine	B	J	O	U	O
The primary structure of protein represents	sequence of amino acid	Helical strucutre	3 D structure	Sub unit of protein	sequence of amino acid
Enzymes are	proteins	carbohydrates	nucleic acids	DNA molecule	Proteins
The first protein squenced by F.Sanger is	Haemoglobin	myoglobin	Insulin	myosin	Insulin
Myoglobin is a	Protein with primary structure	Protein with secondary structure	Protein with tertiary structure	Protein with quartenary structure	Protein with tertiary structure
Haemoglobin has	primary	secondary	tertiary	quarternery structure	quarternery structure
The 3 D structure of protein can be determined by	NMR	X Ray	Spectroscopy	PAGE	X-Ray
SMTP was published in the year	1980	1982	1985	1990	1982
ODBC means	open database connectivity	open direct byte connectivity	open district base connectivity	open dense base connectivity	open database connectivity
Transmission speed range (million bits per second) of coaxial cable ranges from	200 to 500	100 to 200	300 to 600	500 to 1000	200 to 500

UTP means	unshielded twisted pair	unit twisted pair	uniform twisted pair	un twisted pair	unshielded twisted pair
STP means	Shielded twisted pair	Shared twisted pair	Smart twisted pair	Small twisted pair	Shielded twisted pair
The speed of optical fiber than twisted pair	100 times	1000 times	1500 times	2000 times	1000 times
The actual machinery in a computer is called the	machinery	hardware	software	fleshware	machinery
The major components in computers are	joystick	CPU	pendrive	Bluetooth	CPU
IBM S/390 is a	microcomputer	laptop	mainframe	supercomputer	mainframe
Modern web browers use	XHTML	HTTL	MTML	NTML	XHTML
Computer network that spans a city	Metropolitan area network	Mail area network	Messenger area network	Merge area network	Metropolitan area network
Speed of transmission in optical fiber cable	trillion bits/sec	million bits/sec	billion bits/sec	Giga bits/sec	trillion bits/sec
Communication satellites present in space, about	20000 miles above equator	22000 miles above equator	10000 miles above equator	15000 miles above equator	22000 miles above equator
Which is the part that transmits the data from one part of the computer to another?	Bus	CPU	hard disk	software	BUS
OAN means	Open area network	Object area network	Oriented area network	Office area network	Office area network
VPN means	Virtual private network	Vertical position network	Visible private network	Vertical spectrum network	Virtual private network
Primary storage is	software	hardware	external	Internal	hardware
Speed of Wireless LAN	upto 100 Mbs/s	upto 100 kbs/s	upto 1 Gbs/s	upto 10 Mbs/s	upto 1Gbs/s
PAN means	Private area network	Personnel area network	Public area network	Private access network	Personnel area network
HTTP means	Hypertext transfer protocol	High transfer protocol	Horizontal transfer protocol	High transmit protocol	Hypertext transfer protocol
URI means	Universal resource information	Uniform Resource Identifier	Uniform twisted Identifier	Useage resource identifier	Uniform Resource Identifier
Modern web browers use	XHTML	HTTL	MTML	NTML	XHTML
The hardware component that is responsible for executing the instructions is called as	CPU	IPU	PPU	RPU	CPU
Primary storage also known as	Main memory	storage memory	Secondary memory	last memory	Main memory

The place where the programs and data are stored temporarily during processing is called as	Primary storage	Secondary Storage	Tertiary storage	Network storage	Primary storage
IRC stands for	Internet Relay Chat	Internet relative choice	Intranet Relay chat	Intranet relative choice	Internet Relay Chat
Emails, Usenet news, IRC are referred as	Internet suite	Intranet suite	Internet source	Intranet source	Internet suite
FTP means	File transfer protocol	File transfer process	File transmit program	File divert protocol	File transfer protocol
System utilities and other operating services are provides by	System support software	System support hardware	System support service	System surround service	System support software
LAN stands for	Local Area Network	Local internet network	Local wide network	Local access news	Local Area Network
WAN stands for	Wide Area Network	Local internet network	Local wide network	Local access news	Wide Area Network
URI stands for	Uniform Resource Identifier	Unified Resource Identifier	Unidentified Resource Identifier	Universal Resource Identity	Uniform Resource Identifier
URIs retrieved by	HTTP	HIIP	CPU	HTML	HTTP

The computerized language used to store and organize data is called as	Database	Databasic	Data store	Warehouse	Database
Databases are compose of	Hardware and software	Information	Value	Knowledge	Information
Databases enable the easy retrieval of	Information	Internal	External	Resource	Information
The original biological data containing database is called as	Primary database	Secondary database	Tertiary database	Structure database	Primary database
Expansion of EMBL is	European Molecular Biology Laboratory	DNA Data Bank of Japan	INSDC	Protein information resource	European Molecular Biology Laboratory
DDBJ means	DNA Data Bank of Japan	European Molecular Biology Laboratory	INSDC	Protein information resource	DNA Data Bank of Japan
Three primary database together constitutes	INSDC	SWISS-PROT	TrEMBL	PIR	INSDC
INSDC stands for	International Nucleotide Sequence Database Collaboration	International Protein Sequence Database Collaboration	International Nucleotide Structure Database Collaboration	Internal protein Sequence Database Collaboration	International Nucleotide Sequence Database Collaboration
The secondary database which provides sequence annotation is	SWISS-PROT	TrEMBL	PIR	INSDC	SWISS-PROT
TrEMBL contains	translated nucleic acid database	truncated nucleic acid database	translated protein database	truncated protein database	translated nucleic acid database
PIR stands for	Protein information resource	SWISS-PROT	TrEMBL	Primary Internal resource	Protein information resource
The primary source of PROSITE is	SWISS-PROT	UNI-PROT	OWL	PROSITE	SWISS-PROT
The primary source of Profile is	SWISS-PROT	UNI-PROT	OWL	PROSITE	SWISS-PROT
The primary source of PRINTS is	OWL	UNI-PROT	SWISS-PROT	PROSITE	OWL
The primary source of BLOCKS is	PROSITE	UNI-PROT	OWL	BLOCKS	PROSITE
The primary source of IDENTIFY is	BLOCKS	UNI-PROT	OWL	PROSITE	BLOCKS

The database maintains the abstracts of agriculture and parasitic diseases is	CAB International	MEDLINE	AGRICOLA	EMBASE	CAB International
Agricultural field bibliographic database is	AGRICOLA	MEDLINE	BIOSIS	EMBASE	AGRICOLA
Medical field bibliographic database is	MEDLINE	BIOSIS	AGRICOLA	EMBASE	MEDLINE
Publically available bibliographic database is	PUBMED	MEDLINE	AGRICOLA	EMBASE	PUBMED
The database contains the informations of specific organisms is	Organism specific database	Bibliographic database	Biology database	Protein database	Organism specific database
TAIR database contains the informations of	Arabidopsis	Rice	Maize	Sorghum	Arabidopsis
The links for NCBI is _____	WWW.ncbi.co.in	www.ncbi.nih.gov	www.ncbi.com	www.ncbi.nih.gov	www.ncbi.nih.gov
The links for EMBL is _____	WWW. Embl.com	WWW.ebi.uk	WWW.ebi-ac.uk/embl	http://WWW.embl.ac.uk	WWW.ebi-ac.uk/embl
The links for DDBJ is _____	WWW.ddbj.nig.ac.jp	www.ddbj.com	www.ddbj.ac.in	www.ddbj.nlm.com	WWW.ddbj.nig.ac.jp
The link for SWISS PROT is _____	www.swissprot.com	www.expansy.org.ncbi	www.swiss.ac.in	www.expansy.com	www.expansy.org.ncbi
Print database is otherwise known as _____	Nucleotide database	Pattern Database	protein database	Structural database	Pattern Database
The software programs for easy access and retrieval of data is called as	Database management system	Knowledge management system	Database maintainence service	Data managing system	Database management system
Type of database uses set of tables to organize the data is called as	Relational database	Operational database	Object database	Resource databae	Relational database
The language programm used to create relational database called as	Structured query language	Sequence query language	Secondary quary language	Stored query language	Structured query language

Type of database developed to store data as object is called as	Object-oriented database	Relational database	Operational database	Oriental database	Object-oriented database
The language program used to create object-oriented database is	C++	UNIX	WINDOWS	EXCEL	C++
PDB stands for	Protein Data Bank	Structure database	Nucleic acid Data Bank	Cambridge Crystallographic Data Centre	Protein Data Bank
NDB stands for	Nucleic acid Data Bank	Protein Data Bank	Structure database	Cambridge Crystallographic Data Centre	Nucleic acid Data Bank
The database of small molecule structure	CCDC	NDB	PDB	EMBASE	CCDC
CCDC stands for	Cambridge Crystallographic Data Centre	Nucleic acid Data Bank	Protein Data Bank	Structure database	Cambridge Crystallographic Data Centre
CCDC database contains	Protein-Ligand Interactions	Protein-Protein Interactions	Protein-Lipid Interactions	Protein-DNA Interactions	Protein-Ligand Interactions
SWISS-PROT was established by	University of Geneva	University of Florida	University of Singapore	University of Delhi	University of Geneva
SWISS-PROT was maintained by	SIB	PIR	NCBI	EMBL	SIB
SWISS-PROT was established during	1986	1996	2006	2011	1986
NCBI's stand alone sequence submission software is	Sequin	BankIt	tbl2asn	Barcode submission tool	Sequin
The database combines	SWISS-PROT, TrEMBL, PIR	EMBL	NCBI	EBI	SWISS-PROT, TrEMBL, PIR
GSS stands for	Genome Survey Sequences	Barcode submission tool	Expressed sequence tags	Protein information resource	Genome Survey Sequences
Worldwide web is called as	Hypermedia database	External database	End-user database	Distributed database	Hypermedia database
RDBMS includes	Data definition language	Dot definition language	Duty definition language	Data decision language	Data definition language
RDBMS stands for	Relational DBMS	Object DBMS	Oriented DBMS	Other DBMS	Relational DBMS

ODBMS stands for	Object DBMS	Oriented DBMS	Other DBMS	Official DBMS	Object DBMS
Rice genome sequences released at	1990	1995	2000	2005	2005
EMBL created at	1990	1992	1994	1996	1994
Commercial product for the medical literature is	EMBASE	BIOSIS	AGRICOLA	MEDLINE	EMBASE
The inheritor of the old Biological Abstracts database is	BIOSIS	MEDLINE	AGRICOLA	EMBASE	BIOSIS
PDB is the major respiratory of ---------- stuctures	DNA	RNA	Protein	cDNA	Protein
Which electrophoresis used in proteome databases	1D gel	2D gel	3Dgel	4 D gel	2 D
Protein sequence determines ……	genetic variation	genetic disorders	protein structure	Sequence similarity	Protein structure
Taxonomy database deals with	Genetic sequence	Protein structure	DNA structure	Taxonomy of organisms	Taxonomy of organisms
HIV database deals with	HIV virus	STD	Virus	infectious diseases	HIV virus
GOLD is a	Protein database	Genome database	Domain	Motif	Genome database
Which of the following is a characterstics of database	Expensive	Complexity	unidentity	Specificity	Specificity
Proteomics is the study of	study of mRNA	study of cDNA	study of siRNA	study of proteins	study of proteins
Genomics is the study of	DNA c	RNA	Gene	Protein	Gene
The database contains experimentally determined macromolecular structures is	Structure database	PDB	NDB	CCDC	Structure database

Example of progresive method of multiple sequence alignment is	Clustal	Iterative approach	Exhaustive alignment	Text based	Clustal
Reptitive refinement of sub-optimal alignment is	Iterative approach	Exhaustive alignment	Text based	Heuristic	Iterative approach
The percentage matches of similar physiochemical characters of amino acids between two aligned sequences is called as	Sequence similarity	Sequence identity	Global alignment	Local alignment	Sequence similarity
If two aligned sequences assumes similar over their entire length is called as	Global alignment	Sequence similarity	Sequence identity	Local alignment	Global alignment
Sequence similarity is a statement of	Quantitative	qualitative	relative	average	Quantitative
Sequence similarity level depends on the sequence	Type and length	Type and average	length	Type	Type and length
Molecular fossils contains the information of	DNA and proteins	DNA and RNA	DNA and Lipids	Lipids and proteins	DNA and proteins
No sampling bias involved in	Molecular fossil	Phylogenetics	Clade	Morphological	Molecular fossil
The branching pattern in a tree is called as	Tree topology	Root	Node	Branches	Tree topology
If all branches bifurcate in phylogenetic tree is referred as	Dichotomy	Root	Node	Branches	Dichotomy
Neighbor joining method builds a tree based on the method of	Stepwise distance matrix	Phylogram	Cladogram	Polytomy	Stepwise distance matrix
The statistical technique that tests the sampling errors of a phylogenetic tree is called as	Bootstrapping	Phylogram	Cladogram	Polytomy	Bootstrapping
Process of aligning two sequence is called as	Pairwise alignment	Protein alignment	Multiple alignment	Process alignment	Pairwise alignment
The products of evolution are	DNA and proteins	DNA and RNA	DNA and Lipids	RNA and Lipids	DNA and proteins
The chance of being identical unrelated nucleotide sequence is	25%	30%	40%	50%	25%

The chance of being identical unrelaed protein sequence is	5%	10%	15%	20%	5%
If two aligned sequences assumes similar at local region is called as	Local alignment	Global alignment	Sequence similarity	Sequence identity	Local alignment
Graphical way of comparing two sequences in two dimensional matrix is	Dot matrix	Dot matcher	Dottup	Dot analysis	Dot matrix
Simultaneously analyzing all possible aligned positions is called as	Exhaustive alignment	Iterative approach	Text based	Heuristic	Exhaustive alignment
Clustal W version provides the interface of	Text based	Exhaustive alignment	Iterative approach	Heuristic	Text based
The study of evolutionary history of genes and other macromolecules is called as	Molecular phylogenetics	Phylogenetics	Clade	Branches	Molecular phylogenetics
The lines in the phylogenetic tree are called as	Branches	Node	Root	Clade	Branches
The branch point with more than two descendents is called as	Multifurcating node	Root	Node	Branches	Multifurcating node
The phylogeny with multifurcating branches is called as	Polytomy	Root	Node	Branches	Polytomy
Expansion of UPGMA	Unweighted Pair group method using arithmetic average	Phylogram	Cladogram	Polytomy	Unweighted Pair group method using arithmetic average
UPGMA builds a tree based on the method of	Sequential clustering	Phylogram	Cladogram	Polytomy	Sequential clustering
Two taxa shares a unique common ancestor are called as	Sister taxa	Root	Node	Branches	Sister taxa
Number of taxa shares more than one common ancestors are called as	Paraphyletic	Root	Node	Branches	Paraphyletic
The tree branching pattern of evolutionary divergence is called as	Phyogeny	Molecular fossil	Clade	Morphological	Phyogeny

Fossil records contains the information of	Morphological	Biochemical	Chemical	Molecular	Morphological
The scoring function for multiple sequence alignmnet is based on the concept of	Sum of pairs	Squre of pairs	Multiple of pairs	Division of pairs	Sum of pairs
Commonly used algorithms in optimal alignment is	Heuristic	Iterative approach	Exhaustive alignment	Text based	Heuristic
Midnight zone of sequence homology is	< 20 %	< 30 %	< 40 %	< 50 %	< 20 %
The percentage matches of same amino acids between two aligned sequences is called as	Sequence identity	Sequence similarity	Global alignment	Local alignment	Sequence identity
Sequence homology is a statement of	qualitative	quantitative	relative	average	qualitative
If two sequences shares similar physiochemical properties are known as	Sequence similarity	Sequence homology	Sequence difference	Structure homology	Sequence similarity
Degree of sequence conservation in the alignment reveals the evolutionary relatedness of	Different sequence	same sequence	Different segment	Same segment	Different sequence
The basis for structure and function prediction of uncharacterized sequence is	Sequence alignment	Structure alignment	Segment alignment	Sequence analysis	Sequence alignment
Lower similarity depends on the sequence length of	Longer	Shorter	Medium	Average	Longer
Higher similarity depends on the sequence length of	Shorter	Longer	Medium	Average	Shorter
The web-server aligns two sequence in FASTA format is	Dot matcher	Dot matrix	Dottup	Dot analysis	Dot matcher
The web-server aligns two sequence in word format is	Dottup	Dot matcher	Dot matrix	Dot analysis	Dottup
Clustal X version provides the interface of	Graphical	Text based	Exhaustive alignment	Iterative approach	Graphical

The matrix used by clustal for closely related sequences is	BLOSUM62	BLOSUM72	BLOSUM90	BLOSUM250	BLOSUM62
Taxa represents	Present day species	Past day species	Future day species	Average day species	Present day species
The ancestor of extant taxa represented by	Node	Root	Clade	Branches	Node
Unscaled branch length representation of phylogentic tree is called as	Cladogram	Root	Node	Branches	Cladogram
Scaled branch length representation of phylogenetic tree is called as	Phylogram	Root	Node	Branches	Phylogram
Distance based method of phylogentic tree is based on	Distance	Root	Node	Branches	Distance
Character based method of phylogenetic tree is based on	Character	Root	Node	Branches	Character
Common ancestor of all members of the tree represented by	Root	Node	Clade	Branches	Root
Group of taxa descended from a single common ancestor is called as	Clade	Root	Node	Branches	Clade
The matrix used by clustal for divergent sequences is	BLOSUM45	BLOSUM72	BLOSUM90	BLOSUM250	BLOSUM45
The study of the evolutionary history of living organisms using tree like diagrams is called as	Phylogenetics	Molecular fossil	Phyogeny	Clade	Phylogenetics
The handling capacity of SIM scoring matrices is	Tens of Kb	Thousands of Kb	Millions of Kb	Billions of Kb	Tens of Kb
Alignment of multiple related sequence is called as	Multiple sequence alignment	Pair-wise sequence alignment	Multiple structure alignment	Pairwise structure alignment	Multiple sequence alignment
Safe zone of homology between two protein sequence is	30%	40%	50%	60%	30%
Twilight zone of homology between two protein sequence is	20 to 30 %	30 to 40 %	40 to 50 %	40 to 60 %	20 to 30 %

Sequence alignment provides the inference of the two sequence	Relatedness	Difference	Variation	Homology	Relatedness
If two sequence shares high degree of similarity is known as	Sequence homology	Sequence difference	Structure homology	Sequence variance	Sequence homology

Well-conserved regions in multiple sequence alignments	reflect areas of structural importance	reflect areas of functional importance.	reflect areas of both functional and structural importance	reflect areas of both functional and structural importance	
Which server is used to compare three-dimensional protein	DALI	FSSP	SCOP.	CATH.	DALI
Which one of the following tool is used to predict the three-dimensional structure of a	AutoDock	Gromacs	ChemSketch.	Modeller.	Modeller.
Which one of the following tool is not used to predict the three-dimensional structure of a	GLIDE	Swiss-PDB Viewer	JACKAL	Modeller.	GLIDE
Which one of the following tool uses comparative modeling method to predict the three-dimensional structure of a protein	Rosetta.	Threader	CASP	Modeller	Modeller.
Homology modeling is also called as _____.	comparative modeling	abinitio prediction	threading	surface modeling.	comparative modeling
Which one of the following is a computational method to predict the three-dimensional structure of the protein?	X-ray crystallography	NMR	UV Spectroscopy	Threading.	Threading.
Which one of the following is an experimental method to determine the three-dimensional structure of the protein?	Threading	X-ray crystallography	Homology modeling	Abinitio method.	X-ray crystallography
Which method is used for predicting protein tertiary structure in the absence of homology to a known	Comparative modeling	Abinitio prediction	Threading	Surface modeling	Abinitio prediction
Who is the father of Genomics?	Altschul.	Gregor Mendel.	. Richard	Craig Venter	Craig Venter
Which of the following is the character based method?	UPGMA	Maximum Parsimony and Maximum Likelihood.	Maximum Likelihood and Neighbor-Joining	Neighbor-Joining	Maximum Parsimony and Maximum Likelihood.
A phylogenetic tree that explicitly represents number of character changes through its branch lengths	dendogram	cladogram	phylogram	. chronogram	phylogram
Which one of the following is a command based offline tool for molecular structural visualization?	Swiss-PDB Viewer	. RasMol.	QMol.	PyMol.	. RasMol.
Molecular phylogeny can be performed with _____ sequences.	only DNA	only RNA	only Protein	DNA, RNA and Protein	DNA, RNA and Protein
Which one of the following is actually based on MolView?	Raswin.	. QMol.	RasMol	Moldraw.	Moldraw.
Which tool can be used for viewing	Chimera.	QMol.	Arguslab.	ChemSketch.	QMol.
. Energy minimization of a modeled protein can be done using _____.	ChemSketch.	Moldraw	RasMol.	Swiss-PDB Viewer	Swiss-PDB Viewer
Homology modeling can be done using _____.	Swiss-PDB Viewer	QMol	Raswin	Babel.	Swiss-PDB Viewer

Which one of the following tools can be used for both modeling the protein and structure visualization?	Swiss-PDB Viewer	QMol.	RasMol.	ChemSketch.	Swiss-PDB Viewer
Which one of the following tool can be used to generate neighbor joining trees with or without bootstrap	ClustalX	BLAST	Swiss-PDB viewer	ChemSketch	ChemSketch
Which one of the following is more weighted mutation?	Transitions	. Transversions	Transitions and transversion	Deletion	Transitions
Single substitution in the nucleotide sequence is called _____.	single substitution	simple substitution	single nucleotide polymorphism	simple nucleotide polymorphism	single nucleotide polymorphism
Expand UPGMA	Unweighted Pair Group Method with Arithmetic Mean.	Unweighted Pair Group Method with All Mean	Upregulated Gene Method with Arithmetic Mean	Unregulated Genome Method with All Mean.	Unweighted Pair Group Method with Arithmetic Mean.
The study of evolutionary relationships is _____.	Phylogenics	Molecular Evolution.	Cladogenesis.	Cladistics.	Phylogenics
PAUP stands for _____.	. Phylogenetic Analysis Using	Phylogenetic Analysis Using	Proteomic Analysis Using	Phylogenetic Analysis Using	. Phylogenetic Analysis Using
Which one of the following is not a character-based method in tree construction?	Maximum parsimony.	Minimum likelihood	Minimum evolution method.	Neighbor joining.	Neighbor joining.
Which of these methods is a	Unweighted pair group method with	Jukes-Cantor	Minimum evolution.	Maximum parsimony	Unweighted pair group method with
Which is the only method that	Maximum parsimony.	Maximum likelihood	Neighbor Joining	Unweighted pair group method with	Unweighted pair group method with
Template based protein modeling techniques is called as _____	comparative modeling.	surface modeling	threading	abinitio prediction.	comparative modeling.
Which one of the following helps to calculate a structural similarity measure between pairs of structures of protein chains taken from the	CATH.	SCOP	FSSP	DALI.	DALI.
DDD stands for _____.	. Dali Domain Dictionary.	Distance Matrix Alignment Server.	Distance Matrix Domain Dictionary	Distance Domain Dictionary	Dali Domain Dictionary.
A _____ is defined in SCOP as a collection of superfamilies	primary structure of protein	secondary structure of protein.	protein fold	mutated protein sequences	protein fold
SCOP stands for _____.	Similar Classification of Proteins.	Structural Classification of Proteins	Similar Characterization of Proteins	Similar Classification of Proteins.	Structural Classification of Proteins
Which one of the following method predicts the protein structure based on fold recognition?	Comparative modeling.	Threading	Abinitio.	Homology modeling.	Threading
Which server is used to deposit the protein structures in PDB?	ClustalW.	. ClustalX.	ExPASy.	ADIT	ADIT
Which experimental structures cannot be deposited in PDB?	X-ray crystallography.	NMR.	Mass spectrometry.	Comparative modeling	Comparative modeling
PDBID is a combination of _____ number of letters	1	2	3	4	4
PDBID is a _____representation	. SMILES.	ROSDAL.	WLN.	ALPHANUMERIC.	ALPHANUMERIC.
Which of the following is the distance based method?	PGMA	Maximum parsimony.	Maximum likelihood.	Neighbor-Joining	Neighbor-Joining

How many methods are there to predict 3-dimensional structure of a protein?	1	3	5	7	3
Which is a repository for the 3-dimensional structure data for large biomolecules?	NCBI.	EMBL.	Swiss-Prot.	PDB	PDB
Coordinates for known protein structures are housed in?	CATH	SCOP.	. PDBsum.	PDB	PDB
Hydropathy plots are usually used to	beta secondary structure.	transmembrane domains.	alpha secondary structure.	tertiary structure	transmembrane domains.
A term used to classify protein domains according to their secondary structural content and organization is _____.	class	architecture	taxonomy.	homologs.	class
In pairwise alignment result,	true positive.	. true negative	false positive.	false negative.	false positive.
In pairwise alignment result, sequences reported as related homologous represents _____	true positive	true negative.	false negative.	false positive.	true positive
Which one of the PAM matrix	PAM1.	PAM250	PAM60	PAM45	PAM250
Which matrix is based upon the	PAM 1	PAM250	PAM60	PAM 250	PAM 1
Which matrix uses the data on accepted mutations and the probabilities of occurrence of each amino acid to generate a mutation probability?	PAM250	PAM 1	PAM 45	PAM 60	PAM 1
Which method of multiple sequence alignment uses genetic	Progressive	Dynamic Programming.	Genetic Algorithm.	Hidden Markov Model.	Genetic Algorithm.
Two principal ways to construct guide tree in progressive alignment is_____.	UPGMA and Neighbor joining method.	Maximum Parsimony.	Maximum Likelihood.	. all the above.	UPGMA and Neighbor joining method
BLAST2 compares _____ number of sequences.	two	three	four	five	two
Single dot in the sequence	identity.	semi-conserved substitutions.	conserved substitutions.	gaps.	semi-conserved substitutions.
The paired dot in the sequence	conserved substitutions.	semi-conserved substitutions.	gaps	identity.	conserved substitutions.
Speciation event results in_	zoologs.	paralogs.	xenologs.	orthologs.	orthologs.
Gene duplication results in _____.	orthologs	paralogs.	xenologs.	zoologs.	paralogs.
PAM matrices are based on _____ of protein evolution.	Needleman and Wunsch	Smith-Waterman.	Dayhoff model.	Markov model.	Dayhoff model.
What is the full form of BLOSUM?	Blocks of Amino Acid Substitution	Basic Amino Acid Substitution	Blocks of Amino Acid Substitution	Basic Amino Acid Substitution	Blocks of Amino Acid Substitution
Define PAM?	Parallel Align Mutation.	Point Altered Mutation	Point Accepted Mutation.	Point Arranged Mutation	Point Accepted Mutation.
The PRINTS database consists of protein finger prints that define families in the _____ databases	SwissProt/TrEMBL	SwissProt/EMBL.	PIR/TrEMBL.	PIR/EMBL.	SwissProt/TrEMBL