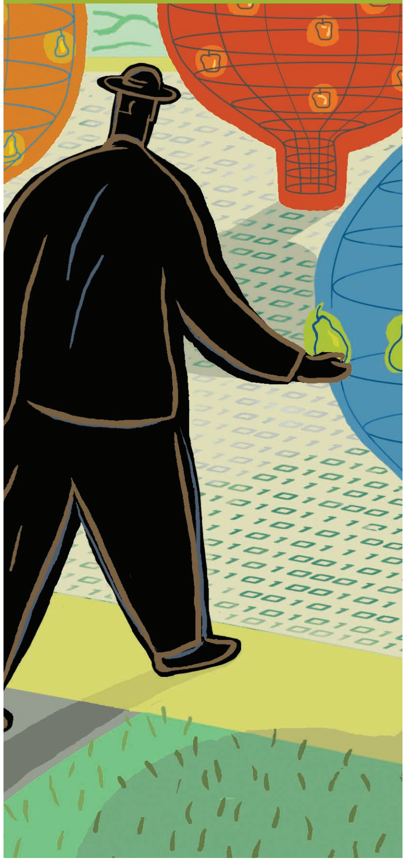

Intelligent Systems in Biology: Why the Excitement?

Richard Lathrop

Biology has rapidly become a data-rich, information-hungry science because of recent massive data generation technologies. Our biological colleagues are designing more clever and informative experiments because of recent advances in molecular science. These experiments and data hold the key to the deepest secrets of biology and medicine, but we cannot fully analyze this data due to the wealth and complexity of the information available. The result is a great need for intelligent systems in biology.

There are many opportunities for intelligent systems to help produce knowledge in biology and medicine. Intelligent systems probably helped design the last drug your doctor prescribed, and they were probably involved in some aspect of the last medical care you received. Intelligent computational analysis of the human genome will drive medicine for at least the next half-century.

Even as you read this, intelligent systems are working on gene expression data to help understand genetic regulation and ultimately the regulated control of all life processes including cancer, regeneration, and aging. Knowledge bases of metabolic pathways and other biological networks make inferences in systems biology that, for example, let a pharmaceutical program target a pathogen pathway that does not exist in humans, resulting in fewer side effects to patients. Modern intelligent analysis of biological sequences today produces the most accurate picture of evolution ever achieved. Knowledge-based empirical approaches currently are the most successful method known for general protein structure prediction, a problem that has been called the Holy Grail of molecular biology. Intelligent literature-access systems exploit a knowledge flow exceeding half a million biomedical articles per year. Machine learning systems exploit heterogenous online databases whose exponential growth mimics Moore's law.

Why now?

So why is this happening now? The answer depends on whether the question is philosophical or practical. Philosophically, it is the inevitable result of the great sweep of intellectual history. Practically, it is because biology is undergoing a data explosion of unprecedented magnitude.

When you look at the intellectual history of the previous century (how strange it seems to term it thus, even now), inevitably you notice that the first half was dominated by chemistry, physics, and mathematics. Quantum mechanics, relativity, and Gödel's incompleteness proof literally changed the mental world in which we live. The second half of the century, however, was dominated by biology and the computing sciences. The genetic code, recombinant organisms, the World Wide Web as an integrated entity, and an intelligent system defeating the world chess champion defined the times. Thus, computational biology sits squarely at the center of the two dominant intellectual forces of the last half-century. Within that historical necessity, the prominent role of intelligent systems is forced on them by the remarkable complexity of the underlying domain.

Biology has become an object of great computational interest because recent technological advances have enabled massive data generation in many critical areas.

1094-7167/01/$10.00 © 2001 IEEE

Both the quantity and diversity of available data are growing rapidly. Figure 1 shows the growth in molecular structures housed in the Protein Data Bank,[1] a repository for 3D biological structure data. Figure 2 shows the growth in DNA sequences housed in GenBank,[2] a repository for 1D nucleotide sequence data. Other major international biological databases are also experiencing rapid growth. Many different high-throughput data generation technologies have come online, providing large amounts of data in diverse areas: combinatorial chemistry for drug discovery, high-throughput screening for bio-assays, two-hybrid protocols for protein interactions, gene expression arrays for monitoring the protein expression of a whole cell, and so on. Add the fact that biomedical research literature contains about 11 million citations and is growing by roughly half a million papers a year, and the amount of data and information to process is staggering. It is exceeded only by the benefits promised in the knowledge we will extract from it.

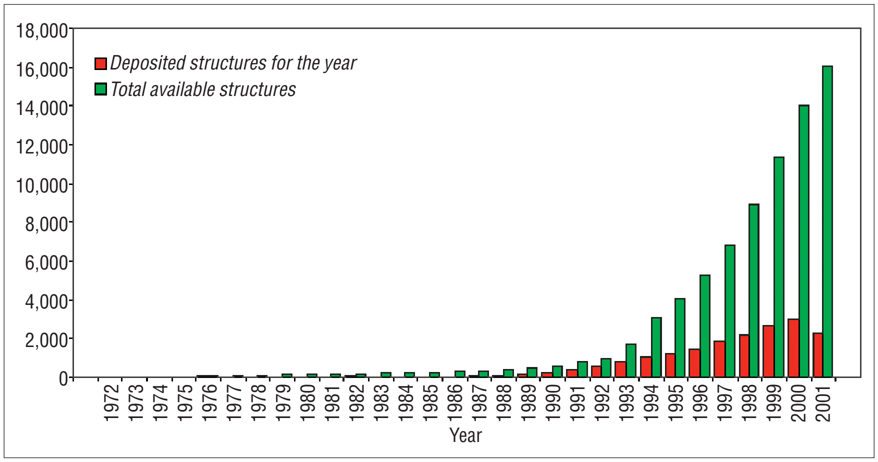

Figure 1. Growth of molecular structure data.

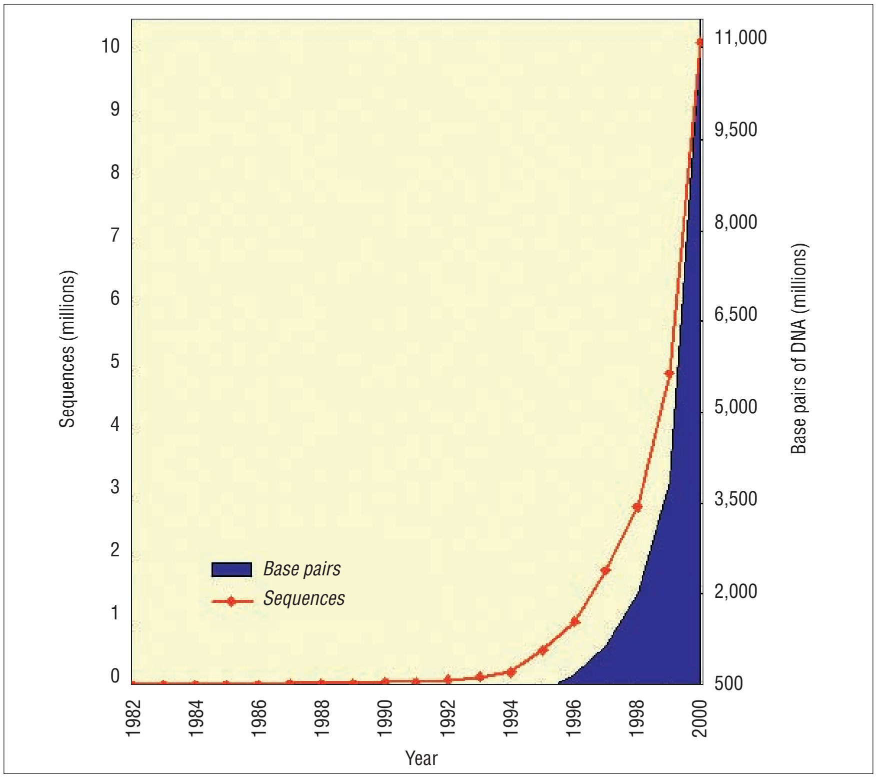

Figure 2. Growth of molecular sequence data.

We can view computer science as a collection of solutions in search of a problem, and the study of life now provides rich problems associated with rich information. The prominent role of intelligent systems arises because, as we all know from personal experience, "Sometimes life just gets complicated!" Intelligent systems are well suited to the complicated domain of biology and medicine. They are robust in the face of inherent complexity, able to extract weak trends and regularities from data, provide models for complex processes, cope with uncertainty and ambiguity, hold the potential to bring content-based retrieval to the biomedical research literature, possess the ontological depth needed to integrate diverse heterogeneous data bases, and in general, aid in the effort to handle semantic complexity with grace.

In this issue

This special issue would have been impossible without the gratifying outpouring of support from the research community involved in intelligent systems for biology. Some 200 scientists have helped produce it: we received 52 manuscripts totaling 163 authors and subjected them to a total of 177 blind reviews by 37 volunteer referees, none of whom was me or an author on any reviewed article. Due to the tremendous volume of high-quality manuscripts, a second special issue in the series is planned for March/April 2002. At all levels, this has been very much a community effort.

The research community behind this special issue is served by a vibrant, and growing, specialized professional society, the International Society for Computational Biology (ISCB), as well as by larger traditional societies such as the IEEE Computer Society, the ACM, AAAI, AAAS, FASEB, the Protein Society, and the Society for Mathematical Biology (I have joined them all, and suggest that you do, too). The ISCB (www.iscb.org) is an excellent contact point for intelligent system practitioners interested in biology.

The current president of the ISCB is Russ Altman, an early champion of intelligent systems in biology[3] and a leading figure in modern bioinformatics. Altman's article opens the "Perspectives" section with an insightful survey titled "Challenges for Intelligent Systems in Biology." This section closes by emphasizing the international character of the field with "The Impact of European Bioinformatics," by Alfonso Valencia, and "The Asia-Pacific Regional Perspective on Bioinformatics," by Satoru Miyano and Shoba Ranganathan. These all provide different views of the field by leading experts.

The articles that follow showcase high points from some of the most interesting and exciting research in the field today. Still, the potential role of intelligent systems is so broad—and the opportunities so great—that this small volume only presents the tip of the iceberg of today's intelligent systems in biology.

"Automatic Pattern Embedding in Protein Structure Models" describes how structural knowledge gained from protein crystals can help predict the structure and function of novel protein sequences. Protein structure prediction from sequence is a central problem of molecular biology. Protein function follows directly from structure and determines the protein's role in biomedical systems. Knowledge-based empirical approaches currently yield the best predictors. The approach here relies on patterns learned from previously seen data and combined according to a Bayesian formulation, which is a familiar architecture in intelligent systems.

"Improving Objectivity and Scalability in Protein Crystallization" brings together robots, machine vision, image analysis, case-based reasoning, and knowledge discovery in a clever and elegant system that targets the rate-limiting step in knowledge acquisition at the atomic level. Almost all of our atom-level knowledge about protein structure and its relation to function comes from x-ray diffraction through protein crystals. It is the necessary first step—getting the protein to crystallize—that most impedes this process. Crystallization often proves difficult or impossible for complex reasons that are poorly understood. This gravely limits our molecular structure knowledge. A system that could learn to produce quick, reliable high-quality protein crystals would revolutionize structural molecular biology.

"Geno2pheno: Interpreting Genotypic HIV Drug Resistance Tests" proposes a machine learning technique for predicting drug resistance in HIV therapy. HIV (through AIDS) is the fourth largest cause of death and the largest cause of productive years of life lost in the developed world and is devastating many developing regions. The article illustrates one of the many medical care settings now touched by intelligent systems. Indeed, the medical domain and medical informatics are long-standing and familiar success stories for AI, and this article continues that fine tradition.

"Toward More Intelligent Annotation Tools: A Prototype" addresses one of the most important problems in bioinformatics: how to extract high-quality information-level summary knowledge from the exponentially growing international scientific databases. This is a rich opportunity for intelligent systems. The article describes how to produce concise descriptions from a protein ID in SwissProt[4] (a repository for 1D protein sequence data). It exploits the database entry annotations that SwissProt already records and so might scale well as SwissProt continues to grow.

"A Knowledge Base for Integrated Biological Systems" uses knowledge representation methods to situate the individual protein functions into an integrated system with multiple interacting players. One ultimate goal of systems biology is to start with knowledge of the complete genome sequence, and thus all proteins encoded by that genome, and proceed automatically to a reconstruction of the biological systems that are implied. The knowledge representation here describes the protein partners and the numerous complex relationships they exhibit. The schema employ concepts long familiar to intelligent systems: classes, associations, hierarchies, an algebraic modelling language, and classification.

"Using Combinatory Categorial Grammar to Extract Biomedical Information" describes research at the intersection of bioinformatics and natural language processing. The archived biomedical literature is a treasure trove of interesting pieces of information, but its huge volume and rapid growth make it increasingly difficult to locate the information most relevant to a particular problem at hand. This article reviews recent approaches to biomedical information extraction, and presents an implemented system that uses a full-fledged natural language grammar.

"Diagnosis Systems in Medicine with Reusable Knowledge Components" looks at medicine from the general viewpoint of knowledge representation and medical informatics, two areas whose fruitful interaction has enriched both AI and medicine. Reusability in knowledge is desirable for much the same reasons it is attractive in software engineering: efficiency, reliability, and economies of scale. If the article's concepts behind reusability of knowledge components scale well and extend to other areas (as hopefully they will), they will help accelerate development of new diagnosis systems in many areas.

Enjoy. ■

Acknowledgments

Two people behind the scenes deserve extra special thanks: Margaret Wyvill kept track of all the manuscripts and reviews, and Mario Espinoza was the Webmaster for the review Web site. Special thanks to Nigel Shadbolt for his vision in initiating this special issue, and also to the *IEEE Intelligent Systems* editorial staff for help and encouragement. Thanks to the PDB and NCBI/GenBank staff for Figures 1 and 2. My deepest debt goes to the volunteer referees, who are responsible for the high scientific quality of these pages and have my shining gratitude and devoted thanks. The following honored colleagues are all hereby named associate guest editors of this special issue: Barb Bryant, Philipp Bucher, Tim Ting Chen, David M. Cooper, Rich Cooper, John Corradi, Terence Critchlow, Steve Culp, Dan Davison, Tom Defay, Francisco M. De La Vega, Valentina Di Francesco, Paolo Frasconi, Frederique Galisson, Richard Goldstein, Harvey Greenberg, Debraj GuhaThakurta, Reece Hart, Dennis Kibler, Mark Lacy, Franz Lang, Gerald Loeffler, Satoru Miyano, Uwe Ohler, Shoba Ranganathan, Isidore Rigoutsos, Paolo Romano, Burkhard Rost, Andrey Rzhetsky, Hershel Safer, Herbert Sauro, Steffen Schulze-Kremer, Vijaya Tirunagaru, Herbert Treutlein, Iosif Vaisman, David Wild, and Tau-Mu Yi.

References

1. H.M. Berman et al., "The Protein Data Bank," *Nucleic Acids Research*, vol. 28, no. 1, 2000, pp. 235–242; www.pdb.org (current 9 Nov. 2001).

2. D.A. Benson et al., "GenBank," *Nucleic Acids Research,* vol. 28, no. 1, 2000, pp. 15–18; www.ncbi.nlm.nih.gov/Genbank (current 9 Nov. 2001).

3. B. Hayes-Roth et al., "Protean: Deriving Protein Structure from Constraints," *Proc. 5th Nat'l Conf. AI* (AAAI 86), AAAI Press, Menlo Park, Calif., 1986.

4. A. Bairoch and R. Apweiler, "The SWISS-PROT Protein Sequence Database and Its Supplement TrEMBL," *Nucleic Acids Research*, vol. 28, no. 1, Jan. 2000, pp. 45–48; http://ca.expasy.org/sprot (current 9 Nov. 2001).

The Author

Richard H. Lathrop is vice-chair of undergraduate education in the Information and Computer Science Department at the University of California, Irvine. In addition to a PhD in artificial intelligence, he holds degrees in electrical engineering, computer science, and mathematics. His research interests include applying intelligent systems and advanced computation to problems in molecular biology, especially protein structure prediction, protein-DNA interactions and genetic regulation, rational drug design and discovery, bio-nanotechnology, and other molecular structure/function relationships. Contact him at ICS Dept. #3425, UCI, Irvine, CA, 92697-3425. Email rickl@uci.edu; www.ics.uci.edu/~rickl.

GLOSSARY

Note: This is a special-purpose glossary compiled from definitions supplied by the authors to explain the meanings of terms as used in their articles. The terms often have wider or other definitions in other contexts. For standard definitons the reader should consult a standard reference text, such as *The Dictionary of Cell and Molecular Biology* (third edition), J.M. Lackie and J.A.T. Dow, eds., Academic Press, London, 1999.

ABC transporter: a kind of transport system whose biological role is both to import substrates into the cell and to export substrates out of the cell.

Align: to put positions of a sequence into correspondence with positions of other sequences. The result is an alignment.

Amino acid (residue): the basic building block of peptides and proteins.

Anamnesis: a patient's case history.

Annotation, gene: (v.) the human and computational activities that lead to the identification, understanding, and storage of auxiliary information about a gene; (n.) the result of such a process.

Base: see nucleotide.

Base pair: a base plus its complement in duplex DNA.

Biological system: a set of biological tangible objects (proteins, protein domains) that are involved in the achievement of a biological process. Systems biology is the study of biological systems.

Chomsky hierarchy: the four classes of formal languages defined by Noam Chomsky as potential models of natural languages.

Classification: (1) categorizing objects according to their characteristics. (2) a top-down process which, given a hierarchy of classes, an object, and a binary relation accounting for to the attachment of an object to a class, looks for the most specific class to which an object can belong.

Collaboratory: defined variously as the use of computing and communication systems in an organized way to support remote collaboration among scientists and others.

Combinatory Categorial Grammar (CCG): adding a limited set of combinators, such as type raising or function composition, to a categorial grammar. This associates each lexical item with its category (or categories) so that the category (or categories) of a given string of lexical items can be computed by combining individual categories with simple function application.

Conserved: preserved during evolution (a property, trait, class, amino acid, base), presumably because it is related to organism fitness.

Cross-resistance: the phenomenon in which a microbe (such as a virus or bacteria) that has acquired resistance to one drug through direct exposure also exhibits resistance to one or more other drugs to which it has not been exposed. Cross-resistance arises because the biological mechanism of resistance to several drugs is the same and is caused by identical genetic mutations.

Crystallization: the process of obtaining suitable single crystals for x-ray structural analysis of a protein.

Cytokines: low-molecular-weight proteins or glycoproteins secreted by white blood cells and various other cells in the body, (assisting in) regulating the development of immune effector cells.

Drug target: a protein that plays a functional role in a disease and is affected by a drug. For HIV-1, for example, all approved drug targets are one or the other of two viral enzymes, namely reverse transcriptase or protease.

Eukaryote: a higher form of organism whose DNA is enclosed by membranes to form a nucleus, comprising plants and animals. Eukaryotes evolved from prokaryotes.

Entrez: the search and retrieval system that integrates information from databases of the US National Center for Biotechnology Information (NCBI), including nucleotide sequences, protein sequences, macromolecular structures, whole genomes, and Medline through PubMed.

Fingerprint: a group of ungapped motifs, often excised from a multiple alignment of a protein family and used as a characteristic signature of family membership.

Fourier transform: a transformation that decomposes or separates a waveform or function into sinusoids of different frequency which sum to the original waveform. It identifies and distinguishes the different frequency sinusoids and their respective amplitudes.

Finite state machine (FSM): an automaton with a finite set of states and a set of transitions from state to state.

G protein: guanine nucleotide binding (G) protein. They are membrane-associated proteins that couple extracellularly activated integral-membrane receptors to intracellular effectors, such as ion channels and enzymes that vary the concentration of second messenger molecules. G proteins are composed of three subunits (alpha, beta, and gamma) which, in the resting state, associate as a trimer at the inner face of the plasma membrane.

GPCR: G protein-coupled receptor (such as muscarinic receptor, opioid receptor). They constitute a vast protein family that encompasses a wide range of functions (including various autocrine, paracrine, and endocrine processes). They show considerable diversity at the sequence level, on the basis of which they can be separated into distinct groups. One group, the so-called rhodopsin-like GPCRs, represent a widespread protein family that includes hormone, neurotransmitter, and light receptors, all of which transduce extracellular signals through interaction with G proteins. Although their activating ligands vary widely in structure and character, the amino acid sequences of the receptors are very similar and are believed to adopt a common structural framework comprising seven trans-membrane helices.

GO: gene ontology.

High-throughput: environments or processes that can deliver or process large amounts of data with sustainable performance over a period of time.

Homology: a similarity arising from shared evolutionary history. The term dates from 1656. Homology specializes the term *similarity* to require divergence from a common ancestor; it is an error to speak of homology unless both similarity and a common origin are present. The result of evolutionary divergence is witnessed, at the protein level, as shared sequence or structural similarity. *Analogy* is similarity without shared evolution. *Paralogy* is gene duplication within a species. *Orthology* is between species.

HPSG: *h*ead-driven *p*hrase *s*tructure *g*rammar.

IE engine: an automated engine for information extraction.

Image feature extraction: the first stage of image analysis, during which descriptive features are extracted from image data to be used during pattern discovery.

Knowledge discovery: extracting and describing novel and potentially useful patterns in data.

Kringle domain: an autonomous protein structural domain (named after the shape of a Danish pastry) found throughout the blood-clotting and fibrinolytic proteins. They are believed to play a role in binding mediators (such as membranes, other proteins, or phospholipids) and in the regulation of proteolytic activity. The kringle domain is characterised by a triple loop, 3-disulphide bridge structure, whose conformation is defined by a number of hydrogen bonds and small pieces of antiparallel beta-sheet.

Lipocalins: a diverse protein family composed mainly of extracellular proteins that display high specificity for small hydrophobic molecules. Functions of these proteins include transport of nutrients, control of cell regulation, pheromone transport, cryptic colouration and synthesis of prostaglandins.

Medline: a literature database, created and maintained by the US National Library of Medicine, containing articles from over 4,300 biomedical journals.

MeSH: *m*edical *s*ubject *h*eadings. The U.S. National Library of Medicine's controlled vocabulary thesaurus, used to index citations primarily to reflect subject content.

Motif: a pattern or string of letters in a biological sequence (typically protein or DNA sequences) whose general character is repeated, or conserved, at corresponding positions in a multiple alignment of related sequences. Motifs are of interest because they may correspond to structural or functional elements within the sequences they characterize.

MUC: message understanding conference.

Multiple alignment: a horizontal stack of sequences within which gaps (insertions or deletions) are used to bring equivalent parts of the sequences into correct register at corresponding positions.

Mutant: an organism (or protein) that changes one or more bases of DNA (or amino acids of protein) relative to the standard organism of that species, often resulting in an organism (or protein) that is abnormal or defective in some way.

NN: noun (singular or mass), a part-of-speech (pos) tag.

Non-Nucleoside Reverse Transcriptase Inhibitor (NNRTI): a drug that inhibits the action of the HIV-1 reverse transcriptase enzyme, thus blocking viral replication. In contrast to nucleoside reverse transcriptase inhibitors (NRTI), NNRTI works by binding directly to the reverse transcriptase.

Nucleoside Reverse Transcriptase Inhibitor (NRTI): a drug that mimics a nucleoside. These compounds, also named Nucleoside Analogs, suppress retroviral replication by interfering with the reverse transcriptase enzyme. The defective synthetic nucleosides cause premature termination of the viral DNA chain. Phosphorylation convert NRTI (prodrugs) into active agents.

Nucleotide/nucleoside: a building block of DNA or RNA.

Object-based knowledge representation with associations: a subset of knowledge representation paradigms, which includes a well-defined representation language and algorithms over the represented knowledge, where factual and generic knowledge are all described and related by the way of predefined constructs such as class, objects, N-ary associations, and inheritance.

Pattern database: a database that contains information derived from primary sequence data, typically in the form of regular expressions (regular patterns), fingerprints, blocks, profiles, or hidden Markov models. These abstractions represent distillations of the most conserved features of multiple alignments, such that they are able to provide potent discriminators of family membership for newly determined sequences.

Peptide: the generic term for any short string of amino acids; often associated with signaling roles.

Planning: the process of deriving a succession of actions intended to achieve a particular goal.

POS: part of speech. Information that refers to (tags) the specific syntactic categories for the lexical items in a sentence. For instance, "saw" in "John saw Mary" is *pos-tagged* as VBD (verb, past tense), whereas the same in "John picked up the saw" is pos-tagged as NN (noun, singular or mass).

Precision: a performance measure in information extraction that is the ratio of the number of correctly extracted pieces of information to that of extracted pieces of information (also called specificity).

Primary structure (sometimes called primary sequence): the linear order of amino acids comprising a protein. See secondary, tertiary, and quaternary structure.

Prion protein: a small glycoprotein found in high quantity in the brain of animals infected with certain degenerative neurological diseases, such as sheep scrapie and bovine spongiform encephalopathy, and the human dementias Creutzfeldt-Jacob disease and Gerstmann-Straussler syndrome. Prions are encoded in the host genome and are expressed both in normal and infected cells. During infection, however, the prion molecules become altered and polymerise, yielding fibrils of modified prion proteins.

Prokaryote: a simple form of organism whose DNA is not enclosed by membranes to form a nucleus, comprising mainly bacteria and blue-green algae. Prokaryotes are the evolutionary ancestors of eukaryotes.

Protease: an enzyme that cleaves proteins. HIV-1 protease cleaves the large precursor proteins produced from viral RNA into the component parts (such as enzymes and structural proteins) that are then assembled into new viral particles. Protease is essential for the production of infectious new viral particles.

Protease Inhibitor (PI): a drug which blocks the action of the HIV-1 protease enzyme, thereby preventing the production of infectious new viral particles.

Protein domain: a compact, local, semi-independent folding unit, presumed to have arisen via gene fusion and gene duplication events. Domains need not be formed from contiguous regions of an amino acid sequence: they can be discrete entities, joined only be a flexible linking region of the chain; they can have extensive interfaces, sharing many close contacts; and they can exchange chains with domain neighbors. The combination of domains within a protein determines its overall structure and function. See protein module.

Protein family (includes superfamily, domain family): a collection of proteins whose members, at the amino acid sequence level, share a high degree of similarity, and hence whose structures and functions are expected to be similar. A superfamily can contain many different protein families; their structures are likely to be highly similar, but their sequences show a lower degree of overall similarity, and hence their functions are also likely to be different (but nevertheless related). A domain family also can contain many different protein families; here, the families are usually unrelated in terms of structure and function, and the only common feature is confined to a domain (often a module, such as kringle, SH2, SH3, zinc-finger, and so on; see protein domain).

Protein module: an autonomous folding unit, believed to have arisen largely as a result of genetic shuffling mechanisms. Modules are contiguous in sequence and are often used as building blocks to confer a variety of complex functions on the parent protein. They can be thought of as a subset of protein domains. Examples of modules include Kringle domains, which are autonomous structural units found throughout the blood clotting and fibrinolytic proteins; the ubiquitous DNA-binding zinc fingers, which are small self-folding units in which zinc is a crucial structural component; and the WW module (characterised by two conserved tryptophan residues, hence its name), which is found in a number of disparate proteins, including dystrophin, the product encoded by the gene responsible for Duchenne muscular dystrophy. See protein domain.

Quasi-species: a large population of genetically closely related individuals that result from erroneous reproduction of a small number of progenitors. HIV, for example, begins reproducing after entering the body and produces both perfect copies of itself and copies containing errors (mutants). Thus, in time there is not a single virus species in the body but, instead, a large population of mixed viruses, for example, a quasi-species.

Quaternary structure: the structure of subunits within a multichain–subunit protein. See primary, secondary, and tertiary structure.

Recall: a performance measure in information extraction which is the ratio of the number of correctly extracted pieces of information to that of the relevant pieces of information in the documents of interest (also called sensitivity)

Residue: see amino acid.

Retrovirus: a class of viruses that carry their genetic material in the form of RNA and use the reverse transcriptase enzyme to transcribe their RNA into DNA. The retrovirus family includes spumaviruses, oncoviruses (such as HTLV-1), and lentiviruses (such as HIV-1 and HIV-2).

Reverse transcriptase (RT): a viral enzyme that allows a retrovirus to translate its genetic material (in the form of RNA) into DNA, which is then integrated into the chromosomes of the host cell.

RNA: ribonucleic acid. A nucleic acid similar to and transcribed from DNA, used for information transmission, protein synthesis, and other cellular functions.

Secondary structure: regions of local regularity in the fold of a protein chain, including tightly folded structures, such as alpha helices, and extended structures, such as beta strands. See primary, tertiary, and quaternary structure.

Sensitivity: the probability of appearance of a finding for a certain diagnosis. See recall.

Sequence: the linear order of letters in a string, as nucleotides in a nucleic acid molecule or amino acids in a protein molecule. See primary structure.

Sequencing: the process of determining the order of letters in a string.

SH2 domain: a small protein module containing about 100 amino acid residues and found in a wide variety of protein contexts such as in association with catalytic domains of phospholipases and nonreceptor protein tyrosine kinases; within structural proteins; and in a group of small adaptor molecules. In many cases, when an SH2 domain is present so too is an SH3 domain, suggesting that their functions are interrelated.

SH3 domain: a small protein module containing about 50 amino acid residues and found in a variety of of proteins with enzymatic activity; in adaptor proteins that lack catalytic sequences; and in cytoskeletal proteins. SH3 domains are commonly found in proteins that also contain SH2 domains, suggesting that their functions are interrelated. They are believed to act as protein binding modules and are involved in linking signals transmitted from the cell surface by protein tyrosine kinases to downstream effector proteins.

Sign: any objective evidence of the presence of a disease or disorder.

Tertiary structure: the 3D structure of a protein. See primary, secondary, and quaternary structure.

Test: the procedure of submitting a statement to such conditions or operations as will lead to its proof or disproof or to its acceptance or rejection.

Therapy failure: see treatment response.

Transcription: the process of copying DNA to RNA.

Translation: the process of producing a protein specified by RNA.

Transmembrane domain: a region of a protein sequence that traverses a membrane; for alpha helical structures, this requires a span of 20-25 residues.

Treatment response: in the treatment of HIV-infected patients, response to therapy is often measured in terms of changes in viral load. A considerable reduction of viral load (for example, below the limit of detection) is considered a therapy success, while a rebound in viral load is referred to as therapy failure. See also Viral load.

vCJD: variant Creutzfeldt-Jakob disease. Bovine spongiform encephalopathy (BSE), or mad cow disease, is the presumed progenitor of vCJD, which is believed to be the human equivalent of mad cow disease.

Viral load: the quantity of HIV RNA in the blood, expressed as the number of copies per milliliter of blood plasma. Research indicates that viral load is a predictor of the risk of progression to AIDS. The lower the viral load the longer the time to AIDS diagnosis and the longer the survival time. Viral load testing for HIV infection is used to determine when to initiate or change therapy.

Zinc finger: an unusually small, self-folding domain in which zinc is a crucial component of its tertiary structure; all bind one atom of zinc in a tetrahedral array to yield a finger-like projection, which interacts with nucleotides in DNA. Zinc fingers are found in all transcription factors and also in proteins such as RNA and DNA polymerases.