1

Knowledge-based bioinformatics

Eric Karl Neumann

1.1 Introduction

Each day, biomedical researchers discover new insights about our biological knowledge, augmenting by leaps our collective understanding of how our bodies work and why they fail us at times. Today, in one minute we accumulate as much information as we would have from an entire year just three decades ago. Much of it is made available through publishing and databases. However, any group's effective comprehension of this full complement of knowledge is not possible today; the stream of real-time publications and database uploads cannot be parsed and indexed as accessible and application-ready knowledge yet. This has become a major goal for the research community, so that we can utilize the gains made through the all the funded research initiatives. This is what we mean by biomedical knowledge-driven applications (KDAs).

Knowledge is a powerful concept and is central to our scientific pursuits. However, knowledge is a term that too often has been loosely used to help sell an idea or a technology. One group argues that knowledge is a human asset, and that all attempts to digitally capture it are fruitless; another side argues that any specialized database containing curated information is a knowledge system. The label 'knowledge' comes to connote information contained by an agent or system that (we wish) appears to have significant value (enough to be purchased). Although the freedom to use labels and ideas should not be impeded, an agreed

Knowledge-Based Bioinformatics: From Analysis to Interpretation Edited by Gil Alterovitz and Marco Ramoni

use of concepts like knowledge would help align community efforts, rather than obfuscate them. Without this consensus, we will not be able to define and apply principles of knowledge to relevant research and development issues that would serve the public. The definition for knowledge needs to be clear, uncomplicated, and practical:

(1) Some aspects of Knowledge can be digitized, since much of our lives depends on the use of computers and the Internet.

(2) Knowledge is different from data or stored information; it must include context and sufficient embedded semantics so that its relevancy to a problem can be determined.

(3) Information becomes Knowledge when it is applicable to more general problems.

Knowledge is about understanding acquired and annotated (sometimes validated) information in conjunction with the context in which it was originally observed and where it had significance. The basic elements in the content need to be appropriately abstracted (classification) into corresponding concepts (usually existing) so that they can be efficiently reapplied in more general situations. A future medical challenge may deal with different items (humans vs. animals), but nonetheless share some of the situational characteristics and generalized ideas of a previously captured biomedical insight. Finding this piece of knowledge at the right time so that it can be applied to an analogous but distinct situation is what separates knowledge from information. Since this is something humans have been doing by themselves for a long time, we have typically been associating knowledge exclusively with human endeavors and interactions (e.g., 'sticky, local, and contextual,' Prusak and Davenport, 2000).

KDA is essential for both industrial and academic biomedical research; the need to create and apply knowledge effectively is driven by economic incentives and the nature of how the world works together. In industry, the access to public and enterprise knowledge needs to be both available and in a form that allows for seamless combinations of the two sets. Concepts must enable the bridging between different sources, such that the connected union set provides a business advantage over competitors. Academic research is not that different in having internal and external knowledge, but once a novel combination has been found, validated and expounded, the knowledge is then submitted to peer review and published in an open community. Here, rather than supporting business drivers, scientific advancement occurs when researchers strive to be recognized for their contribution of novel and relevant scientific insights. The free and efficient (and sometimes open) flow of knowledge is key in both cases (Neumann and Prusak, 2007).

In preparation for the subsequent discussions, it is worth clarifying what will be meant by data, information, and knowledge. The experimentalists' definition of data will be used for the most part unless otherwise noted, and that is

information measured or generated by experiments. Information will refer to all forms of digitized resources (aka data by other definitions) that can be stored and recalled from a program; it may or may not be structured. Finally, based on the above discussion, knowledge refers to information that can be applied to specific problems, usually separate from the sources and experiments from which they were derived. Knowledge can exist in both humans and digital systems, the former being more flexible to interpretation; the latter relies on the application of formal logic and well-defined semantics.

This chapter begins by providing a review of historical and contemporary knowledge discovery in bioinformatics, ranging from formal reasoning, to knowledge representation, to the issues surrounding common knowledge, and to the capture of new knowledge. Using this initial background as a framework, it then focuses on individual current knowledge discovery applications, organized by the various components and approaches: ontologies, text information extraction, gene expression analysis, pathways, and genotype–phenotype mappings. The chapter finishes by discussing the increasing relevance of the Web and the emerging use of Linked Data (Semantic Web) 'data aggregative' and 'data articulative' approaches. The potential impact of these new technologies on the ongoing pursuit of knowledge discovery in bioinformatics is described, and offered as practical direction for the research community.

1.2 Formal reasoning for bioinformatics

Computationally based knowledge applications originate from AI projects back in the late 1950s that were designed to perform reasoning and inferencing based on forms of first-order logic (FOL). Specifically, inferencing is the processing of available information to draw a conclusion that is either logically plausible (inconclusive support) or logically necessary (fully sufficient and necessary). This typically involves a large set of chained reasoning tasks that attempt to exhaustively infer precise conclusions by looking at all available information and applying specified rules.

Logical reasoning is divided into three main forms: deduction, induction, and abduction. These all involve working with preconditions (antecedents), conclusions (consequents), and the rules that associate these two parts. Each one tries to solve for one of these as unknowns given the other two knowns. Deduction is about solving for the consequent given the antecedent and the rule; induction is about finding the rule that determines the consequent based on the known precondition; and abduction is about determining the precondition based on the conclusions and the rules followed. Abduction is more prone to problems since multiple preconditions can give rise to the same conclusions, and is not as frequently employed; we will therefore focus only on deduction and induction here.

Deduction is what most people are familiar with, and is the basis for syllogisms: 'All men are mortal; Socrates is a man: Therefore Socrates is mortal!' Deductive reasoning requires no further observations; it simply requires applying

rules to information on preconditions. The difficulty is that in order to perform some useful reasoning, one must have a lot of deep knowledge in the form of rules so that one can produce solid conclusions. Mathematics lends itself well here, but attempts to do this in biology are limited to simple problems: 'P53 plays a role in cancer regulation; Gene X affects P53: Therefore Gene X may play a role in a cancer.' The rule may be sound and generalized, but the main shortcoming here is that most people could have performed this kind of inference without invoking a computational reasoner. Evidence is still scant that such reasoning can be usefully applied to areas such as genetics and molecular biology.

Induction is more computationally challenging, but may have more real-world applications. It benefits from having lots of evidence and observations on which to create rules or entailments, which, of course, there is plenty of in research. Induction works on looking for patterns that are consistent, but can be relaxed using statistical significance to allow for imperfect data. For instance, if one regularly observes that most kinases downstream of NF-kB are up-regulated in certain lymphomas, one can propose a rule that specifies this up-regulation relation in these cancers. Induction produces rule statements that have antecedents and consequents. For induction to work effectively one must have (1) sufficient data, including negative facts (when things didn't happen); (2) sufficient associated data (metadata), describing the context and conditions (experimental design) under which the data were created; and (3) a listing of currently known associations which one can use to specifically focus on novel relations and avoid duplication. Induction by itself cannot determine cause and effect, but with sufficient experimental control, one can determine which rules are indeed causal. Indeed, induction can be used to generate hypotheses from previous data in order to design testable experiments.

Induction relies heavily on the available facts present in sources of knowledge. These change with time, and consequently inductive reasoning may yield different results depending on what information has recently been assimilated. In other words, as new facts come to light, new conclusions will arise out of induction, thereby extending knowledge. Indeed, a key reason that standardized databases such as Gene Expression Omnibus (GEO, www.ncbi.nlm.nih.gov/geo/) exist is so we can discover new knowledge by looking across many sets of experimental data, longitudinally and laterally.

Often, reasoning requires one to make 'open world assumptions' (OWAs) of the information (e.g., *Ling-Ling is a panda*), which means that if a relevant statement is missing (*Ling-Ling is human* is absent), it **must** be assumed plausible unless (1) proven false (*Ling-Ling's parents are not human*), (2) shown to be inconsistent (*pandas and humans are disjoint*), or (3) the negation of the statement is provided (*Ling-Ling is not human*). OWAs affect deduction by expanding the potential solution space, since some preconditions are unknown and therefore unbounded (not yet able to be fixed). Hence, a receptor with no discovered ligand should be treated as a potential receptor for many different signaling

processes (ligands are often associated with biological processes). Once a ligand is determined, the signaling consequences of the receptor are narrowed according to the ligand.

With induction, inference under OWAs will usually be incomplete, since a rule cannot be exactly determined if relevant variables are unknown. Hence some partial patterns may be observed, but they will appear to have exceptions to the rule. For example, a drug target for colon cancer may not respond to inhibitors reliably due to regulation escape through an unbeknownst alternative pathway branch. Once such a cross-talk path is uncovered, it becomes obvious to try and inhibit two targets together, one in each pathway, to prevent any regulatory escape (aka *combinatoric therapy*).

Another relevant illustration is the inclusion of Gene Ontology (GO) terms within gene records. Their presence suggests that evidence exists to recommend assigning a role or location to the gene. However, the absence of the attribute 'regulation of cell communication' could signify a few things: (1) the gene has yet to be assessed for involvement in 'regulation of cell communication'; (2) the gene has been briefly reviewed, and no obvious evidence was found; and (3) the gene has been thoroughly assessed by a sufficient inclusionary criteria. Since there is no way to determine, today, what the absence of a term implies, this would suggest that knowledge mining based on presence or absence of GO terms will often be misleading.

OWAs often cannot be automatically applied to relational database management systems (RDBMSs), since the absence of an entry or fact in a record may indeed mean it was measured but not found. A relational database's logical consistency could be improved if it explicitly indicated which facts were *always* measured (i.e., lack of fact implies measured and not observed), and which ones were *sometimes* measured (i.e., if measured, always stated, therefore lack of fact implies not measured). The measurement attribute would need to include this semantic constraint in an accessible metamodel, such as an ontology.

Together, deduction and induction are the basis for most knowledge discovery systems, and can be invoked in a number of ways, including non-formal logic approaches, for example SQL (structured query language) in relational databases, or Bayesian statistical methods. Applying inference effectively to large corpora of knowledge requires careful planning and optimization, since the size of information can easily outpace the computation resources required due to combinatorial explosion. It should be noted that biology is notoriously difficult to generalize completely into rules; for example, the statement 'P is a protein iff P is triplet-encoded by a Gene' is almost always true, but not in the case of gramicidin D, a linear pentadecapeptide that is synthesized *de novo* by a multi-enzyme complex (Kessler *et al*., 2004). The failure of AI, 25 years ago, was in part due to not realizing this kind of real-world logic problem. We hope to have learned our lessons from this episode, and to apply logical reasoning to large sets of bioinformatic information more prudently.

1.3 Knowledge representations

Knowledge Representations (KRs) are essential for the application of reasoning methodologies, providing a precise, formal structure (ontology) to describe instances or individuals, their relations to each other, and their classification into classes or kinds. In addition to these ontological elements, general axioms such as subsumption (class–subclass hierarchies) and property restrictions (e.g., *P has Child C iff P is a Father* $\vee$ *P is a Mother*) can be defined using common elements of logic. The emergence of the OWL Web ontology language from the W3C (World Wide Web Consortium) means that such logic expressions can be defined and applied to information resources (IRs) across the Web, enabling the establishment of KRs that span many sites over the Internet and many kinds of information resources. This is an attractive vision and could generate enormous benefits, but in order for all KRs to work together, there still needs to be coherence and consistency between the ontologies defined (in OWL) and used. Efforts such as the OBO (Open Biomedical Ontologies) Foundry are attempting to do this, but also illustrate how difficult this process is.

In the remainder of this chapter, we will take advantage of a W3C standard format known as N3 (www.w3.org/TeamSubmission/n3/) for describing knowledge representations and factual relations; the triple predicate form 'A B_{rel} C' is to be interpreted as 'Entity A has relation B_{rel} with entity C.' Any term of the form '?B' signifies a named variable that can be anything that makes the predicate true; for example '?g a Gene' means ?g could be any gene, and the double clause '?p a Protein. ?p is_expressed_in Liver' means any protein is expressed in liver. Furthermore, ';' signifies a conjunction between phrases with the same subject but multiple predicates ('?p a Protein ; is_expressed_in Liver' as in the above). Lastly, '[]' brackets are used to specify any entity whose name is unknown (or doesn't matter) but which has relations contained within the brackets: '?p is_expressed_in [a Neural_Tissue; stage Embryonic].' One should recognize that such sets of triples result in the formation of a system of entity nodes related to other entity nodes, better known as a graph.

1.4 Collecting explicit knowledge

A major prerequisite of knowledge-driven approaches is the need to collect and structure digital resources as KRs (a subset of IRs), to be stored in knowledge bases (KBs) and used in knowledge applications. Resources can include digital data, text-mined relations, common axioms (subsumption, transitivity), common knowledge, domain knowledge, specialized rules, and the Web in general. Such resources will often come from Internet-accessible sources, and it is assumed that they can be referenced similarly from different systems. Web accessibility requires the use of common and uniform resource identifiers (URIs) for each entity as well as the source system; the additional restriction of uniqueness is not

as easy to implement, and can be deferred as long as it is possible to determine whether two or more identifiers refer to the same thing (e.g., owl:sameAs).

In biomedical research, recognizing where knowledge comes from is just as important as knowing it. Phenomena in biology cannot be rigorously proven as in mathematics, but rather are supported by layers of hypotheses and combinations of models. Since these are advanced by researchers with different working assumptions and based on evidence that often is local, keeping track of the context surrounding each hypothesis is essential for proper reasoning and knowledge management. Scientists have been working this way for centuries, and much of this has been done through the use of references in publications whenever (hypothetical) claims are compared, corroborated, or refuted. One recent activity that is bridging between the traditional publication model and the emerging KR approach is the SWAN project (Ciccarese *et al*., 2008), which has a strong focus on supporting evidence-based reasoning for the molecular and genetic causes of Alzheimer's disease.

Knowledge provenance is necessary when managing hypotheses as they either acquire additional supporting evidence (accumulating but never conclusive), or are disproved by a single critical fact that comes to light (single point of failure). Modal logic (see below), which allows one to define hypotheses (beliefs) based on partial and open world assumptions (Fagin *et al*., 1995), can dramatically alter a given knowledge base when a new assumption or fact is introduced to the reasoner (or researcher). As we begin to accumulate more hypotheses while at the same time having to review new information, our knowledge base will be subject to major and frequent inference-driven updates. This dependency argues strongly for employing a common and robust provenance framework for both scientific facts and (hypotheses) models. Without this capability, one will never know for sure on what specific arguments or facts a model is based, hence impeding effective Knowledge Discovery (KD). It goes without saying that this capability will need to work on and across the Web.

The biomedical research community has, to a large extent, a vast set of common knowledge that is openly shared. New abstracts and new data are put on public sites daily whenever they are approved or accepted, and many are indexed by search engines and associated with controlled vocabulary (e.g., MeSH). However, this collection is not automatically or easily assimilated into individual applications using knowledge representations, so that researchers cannot compare or infer new findings against their existing knowledge. This barrier to knowledge discovery could be removed by ensuring that new published reports and data are organized following principles of *common knowledge*.

1.5 Representing common knowledge

Common knowledge refers to knowledge that is generally known (and accessible) by everyone in a given community, and which can be formally

described. Common knowledge usually differs from *tacit knowledge* (Prusak and Davenport, 2000) and *common sense*, both of which are virtually impossible to explicitly codify and which require assumptions that are non-deducible[1]. For these reasons we will focus specifically on explicit common knowledge as it applies to bioinformatic applications.

An example of explicit common knowledge is 'all living things require an energy source to live.' More relevant to bioinformaticists is the central dogma of biology which states: 'genes are transcribed into mRNA which translate into proteins; implying protein information cannot flow back to DNA,' or formally:

```
∀ Protein ∃ Gene (Gene transcribes_into mRNA ∧ mRNA
     translates_into
Protein) ⇒ ¬ (Protein reverse_translate Gene).
```

This is a very relevant chunk of common knowledge that not only maps proteins to genes, but even constrains the gene and protein sequences (up to codon ambiguity). In fact, it is so common, that it has been (for many years) hard-wired into most bioinformatic applications. The knowledge is therefore not only common, but pervasive and embedded, to the point where we have no further need to recode this in formal logic. However, this is not the case for more recent insights such as SNP (single nucleotide polymorphism) associations with diseases, where the polymorphism does not alter the codons directly, but the protein is either truncated or spliced differently. Since the set of SNPs is constantly evolving, it is essential to make these available using formal common knowledge. The following (simplified) example captures this at a high level:

```
∀ Genetic_Disease ∃ Gene ∃ Protein ∃ SNP (SNP within
     Gene ∧ Gene
expresses Protein ∧ SNP modifies Protein ∧ SNP associated
Genetic_Disease) ⇒ SNP root_cause_of Genetic_Disease.
```

Most of these relations (protein structure and expression changes) are being curated into databases along with their disease and gene (and sequence) associations. It would be a powerful supplement if such knowledge rules could be available as well to researchers and their applications. An immediate benefit would be to allow for application to extend their functionality without need for software updates by vendors; simply download the new rules based on common understanding to reason with local knowledge.

Due to the vastness of common knowledge around all biomedical domains (including all instances of genes, diseases, and genotypes), it is very difficult to explicitly formalize all of it and place it in a single KB. However, if one considers public data sources as references of knowledge, then the amount of digitally

[1] Tacit knowledge is often related to human habit and know-how, and is not typically encodable and therefore not easily shared; common sense does involve logic but requires assumptions that cannot be formally defined.

encoded knowledge can be quickly and greatly augmented. This does require some mechanism for wrapping these sources with formal logic, for example associating entities with classes. Fortunately, the OWL-RDF (resource description framework) model is a standard that supports this kind of information system wrapping, whereby entities become identified with URIs and can be typed by classes defined in separate OWL documents. Any logical constraints presumed on database content (e.g., no GO process attribute means no evidence found to date for gene) can be explicitly defined using OWL (and other axiomatic descriptions); these would also be publicly accessible from the main source site.

Common knowledge is useful for most forms of reasoning, since it facilitates making connections between specific instances of (local) problems and generalized rules or facts. Novel relations could be deduced on a regular basis from the latest new findings, and deeper patterns induced from increasing numbers of data sets. Many believe that true inference is not possible without the proper encoding of complete common knowledge. Though it will take time to reach this level of common knowledge, it appears that there is interest in heading towards such open knowledge environments (see www.esi-bethesda.com/ncrrworkshops/kebr/index.aspx). If enough benefits are realized in biomedicine along the way, more organized support will emerge to accelerate the process.

The process for establishing common knowledge can be handled by a form of logic known as modal logic (Fagin *et al*., 1995), which allows different agents (or scientists) to be able to reason with each other though they may have different subsets of knowledge at a given time (i.e., each knows only part of the story). The goal here is to somehow make this disjoint knowledge become common to all. Here, common knowledge is (1) knowledge (φ) all members know about ($E_G\varphi$), and importantly (2) something known by all members to be known to the other members. The last item applies to itself as well, forming an infinite chain of 'he knows that she knows that he knows that...' signifying *complete awareness of held knowledge*

$$\left(C\varphi = \lim_{n\to\infty} E^{n\cdots} E^2 E^1_G \varphi\right).$$

Another way to understand this, is that if Amy knows X about something, and Bob knows only Y, and X and Y are both required to solve a research problem (possibly unknown to Amy and Bob), then Amy and Bob need to combine their respective sets as common knowledge to solve a given problem. In the real world this manifests itself as experts (or expert systems) who are called upon when there is a gap in knowledge, such as when an oncologist calls on a bioinformatician to help analyze biomarker results. Automating this knowledge expert process could greatly improve the efficiency for any researcher when trying to deduce if their new experimental findings have uncovered new insights based on current knowledge.

In lieu of a formal method for accessing common knowledge, researchers typically resort to searching through local databases or using Google (discussed later) in hopes of filling their knowledge gaps. However, when searching a

RDBMS with a query, one must know how to pose the query explicitly. This often results in not uncovering any new significant knowledge, since one requires sufficient prior knowledge to enter the right query in the first place, in which case one is searching only 'under the street lamp.' More likely, only particular instances of facts are uncovered, such as dates, numeric attributes and instance qualia. This is a distinguishing feature that separates databases from knowledge bases, and illustrates that databases can support at best very focused and constrained knowledge discovery. Specifically, queries using SQL produce limited knowledge, since they typically do not uncover generalized relations between things. Ontological relations rely on the ability to infer classes, groups of relations, transitivity (multi-joins), and rule satisfiability; these are the instruments by which general and usable knowledge can be uncovered. Standard relational databases (by themselves) are too restrictive for this kind of reasoning and do not properly encode (class and relation) information for practical knowledge discovery.

In many cases, in the bioinformatics community this has come to be viewed as knowledge within databases. For example, the curated protein database Swiss-Prot/UniProt is accepted as a high quality source of reviewed and validated knowledge for proteins, including mutational and splice variants, relations to disorders, and the complexes which they constitute. In fact, it is often the case that curated sets of information are informally raised to the level of knowledge by the community. This definition is more about practice and interpretation than any formal logic definition. Nonetheless, it is relevant and valid for the community of researchers, who often do apply logic constraints on the application of this information: if a novel protein polymorphism not in Swiss-Prot is discovered and validated, it is accepted as real and eventually becomes included into Swiss-Prot.

Nonetheless, databases are full of valuable information and could be re-formatted or wrapped by an ontological layer that would support knowledge inference and discovery, defined here as *implicit knowledge resources* ($IR \rightarrow KR$). If this were to happen, structured data stores could be federated into a system of biomedical common knowledge: knowledge agents could be created that apply modal logic reasoning to crawl across different *knowledge resources* on the Web in search of new insights. Suffice it to say, practical modal logic is still an emerging and incomplete area of research. Currently, humans are best at identifying which new facts should be incorporated into new knowledge regarding a specific subject or phenomenon; hence researchers would be best served by being provided with intelligent *knowledge assistants* that can help identify, review, compare, and assimilate new findings from these biomedical IRs. There is a lot of knowledge in biology that could be formally common, consequently there is a clear need to transform public biomedical information sources to work in concert with knowledge applications and practices. Furthermore, this includes the Web, which has already become a core component of all scientific research communities.

1.6 Capturing novel knowledge

Not everything can be considered common knowledge; there are large collections of local-domain knowledge consisting of works and models (published and pre-published) created by individual research groups. This is usually knowledge that has not been completely vetted or validated yet (hypotheses and beliefs), but nonetheless can be accessed by others who wish to refute or corroborate the proposed hypotheses as part of the scientific method. This knowledge is connected to and relies on the fundamentals of biology, which are themselves common knowledge (since they form the basis of scientific common ground). So this implies that we are looking for a model which allows connecting local knowledge easily with common knowledge.

Research information is knowledge that is in flux; it is comprised of assumptions and proposed models (mechanisms of action). In modal logic (Fagin *et al*., 1995) this is comparable to the KD45 axioms: an agent (individual or system) can believe in something not yet proven true, but if shown to be false, the agent cannot believe in it anymore; that is, logic contradictions are not allowed. KD45 succinctly sums up how the scientific process works with competing hypotheses, and how all parallel hypotheses can co-exist until evidence emerges that proves some to be incorrect.

Therefore, the finding (by a research group), that a mutation in the *BRCA2* gene is always associated with type 2 breast cancer, strongly argues against any other gene being the primary cause for type 2 susceptibility. Findings that have strong causal relations, such as nucleotide level changes and phenotypes of people always carrying these, are prime examples of how new-findings knowledge can work together with common knowledge. As more data is generated, this process will need to be streamlined and automated; and to prevent too many false positives from being retained, the balanced use of logic and statistics will be critical.

The onslaught of large volumes of information being generated by experiments and subsequent analyses requires proper data set tracking, including the capture of experimental conditions for each study. The key to managing all these associated facts is enforcing data provenance. Without context and provenance, most experimental data will be rendered unusable for other researchers, a problem already identified by research agencies (Nature Editorial, 2009). Provenance will ensure a reliable chain of evidence associated by conditions and working hypotheses that can be used to infer high-value knowledge associations from new findings.

1.7 Knowledge discovery applications

Once common and local knowledge are available to systems in a machine-interpretable form, the construction and use of knowledge-discovery applications that can work over these sources becomes practical and empowering. KDA has its

roots in what has been labeled Knowledge Discovery and Data Mining (KDD, Fayyad *et al*., 1996), which consists of several computational approaches that came together in the mid 1990s under a common goal. The main objective of KDD is to turn collected data into knowledge, where knowledge is something of high value that can be applied directly to specific problems. Specifically, it had become apparent that analysts were 'drowning in information, but starving for knowledge' (Naisbitt, 1982). It was hoped that KDD would evolve into a formal process that could be almost entirely computationally driven. It was to assist knowledge workers during exploratory data analysis (EDA) when confronted by large sets of data. The extraction of interesting insights via KDD follows more or less inductive reasoning models.

KDD utilizes approaches such as first-order logic and data mining (DM) to extract patterns from data, but distinguishes itself from DM in that the patterns must be validated, made intelligently interpretable, and applicable to a problem. More formally, KDD defines a process that attempts to find expression patterns (E) in sets of Facts (F) that have a degree of certainty (C) and novelty (N) associated with them, while also being useful (U) and simple (S) enough to be interpreted. Statistics plays a strong role in KDD, since finding patterns with significance requires the proper application and interpretation of statistical theories. Some basic KDD tools include Decision Trees, Classification and Regression, Probabilistic Graphs, and Relational Learning. Most of these can be divided into supervised learning and unsupervised learning. Some utilize propositional logic more than others.

Key issues that KDD was trying to address include:

- Data classification
- Interpretation of outcomes (uncovering relations, extraction of laws)
- Relevance and significance of data patterns
- Identification and removal of confounding effects (Simpson's paradox, http://plato.stanford.edu/entries/paradox-simpson/).

Patterns may be known (or hypothesized) in advance, but KDD is supposed to aid in the extraction of such patterns based on the statistical structure of the data and any available domain knowledge. Clearly, information comes in a few flavors: quantitative and qualitative (symbolic). KDD was intended to take advantage of both wherever possible. Symbolic relations embedded in both empirical data (e.g., *what conditions were different samples subjected to?*) and domain knowledge (e.g., *patient outcomes are affected by their genotypes*) begin to demonstrate the true symbolic nature of information. That is, data is about tying together relations and attributes, whether it is arranged as tables of values or sets of assertions. The question arises, how can we use this to more efficiently find patterns in data? The key here is understanding that relational data can be generalized as data graphs: collections of nodes connected by edges, analogous to how formal relational knowledge structures are to function (see above).

Indeed, all the KDD tools listed have some form of graph representation: decision trees, classification trees, regression weighted nodes, probabilistic graphs, and relational models (Getoor *et al*., 2007). The linked nodes can represent samples, observations, background factors, modeling components (e.g., θ_i), outcomes, dependencies, and hidden variables. It would follow that a common way to represent data and relational properties using graphs could help generalize KDD approaches and allow them to be used in concert with each other. This is substantial, since we now have a generalized way for any application to access and handle knowledge and facts using a common format system, based on graph representation (and serialized by the W3C standard formats, RDF-XML or RDF-N3).

More recently, work by Koller and others has shown that the structure of data models (relations between different tables or sets of things) can be exploited to help identify significant relations within the data (Getoor *et al*., 2007). That is, the data must already be in a graph-knowledge form in order to be effectively mined statistically. To give an example, if a table containing tested-subject responses for a treatment is linked to the treatment-dosing table and the genetic alleles table, then looking for causal response relations is a matter of following these links and calculating the appropriate aggregate statistics. Specifically, if one compares all the responses in conjunction with the drug and dosing used as well as the subject's genotype, then by applying Bayesian inference, strong interactions between both factors can be identified. Hence data graph structures can be viewed as first-order 'hypotheses' for potential interactions.

By the mid 1990's, the notion of publishing useful information on the Web began to take off, allowing it to be linked and accessed by other sites: the Web as a system of common knowledge took root and applications began to work with this. This was followed by efforts to define ontologies in a way that would work from anywhere on the Web, and with anything that was localized anywhere on the Web. A proposal was eventually submitted to the DARPA (Defense Advanced Research Projects Agency) program to support typing things on the Web. It was funded in 2000 and became known as the DAML (DARPA Agent Mark-up Language, www.daml.org/) project. This became the forerunner of the Semantic Web, and eventually transformed into the OWL ontology language that is based on Description Logic (DL).

Dozens of applications for KDD have been proposed in many different domains, but its effectiveness in any one area over the other is unclear. To this end, an open challenge has been initiated, called the KDDCup (www.kdnuggets.com/datasets/kddcup.html), to see how well KDD can be applied to different problem spaces. It has gained a large following in bioinformatics, addressing such diverse areas as:

- Prediction of gene/protein function and localization
- Prediction of molecular bioactivity for drug design
- Information extraction from biomedical articles
- Yeast gene regulation prediction

- Identification of pulmonary embolisms from three-dimensional computed tomography data
- Computer Aided Detection (CAD) of early stage breast cancer from X-ray images.

KDD was conceived during a time when the shortcomings of AI had surfaced, and the Web's potential was just emerging. We are now in an age where documents can be linked by other documents anywhere in the world; where communities can share knowledge and experiences; where data can be linked to meaning. The most recent rendition of this progress is the Semantic Web.

1.8 Semantic harmonization: the power and limitation of ontologies

One of the most important requirements for data integration or aggregation is for all data producers and consumers to utilize common semantics (Rubin *et al.*, 2006). In the past, it had been assumed that data integration was about common formats (syntax), but that assumed that if one *knows* the structure, one can *infer* the data meaning (semantics). This is now known to be grossly oversimplified, and semantics must also be clearly defined. RDF addressed the syntax issue by forcing all data relations to be binary based, therefore modeling all components as triples (subject, relation/property, object).

The emergence of the W3C OWL ontology standard has enabled the formal definition of many biological and medical concepts and relations. OWL is based on description logic, a FOL formalism that was developed in the 1980s to support class (concept) subsumption and relations between instances of classes. Using OWL, a knowledge engineer can create class hierarchies of concepts that map to real-world observations; for instance, 'Genes are encoded in DNA and themselves encode proteins.' OWL's other key feature is that it can be referenced from anywhere on the Web (e.g., used by a database) and incorporated into other non-local logical structures (ontology extension). It was designed to so that any defined ontological components are identifiable nodes on the Web; that is, all users can refer to the same defined Class. The most current version of OWL is OWL2, based on SROIQ logic supporting more expressive logic (Horrocks *et al.*, 2006). The OWL format is modeled after the Resource Description Framework (RDF) that will be described later.

The OWL standard allows knowledge systems to utilize ontologies defined by various groups, such as Gene Ontology, UniProt, BioPAX, and Disease Ontology. Data sets that one wishes to align with the ontologies now can apply a well-specified mechanism: simply reference the ontology URI from within a data documents and system. By doing so, all the data in the system is formally associated with concepts, as well as the relations concepts have with each other. Any third party also looking at the data can instantly find (over the Web) which ontologies were used to define the set.

Many of the current activities around developing ontologies in OWL are about defining common sets of concepts and relations for molecular biology (genes, proteins, and mechanisms) and biomedicine (diseases and symptoms). However, there is still no general agreement of how (completely) to define basic concepts (e.g., gene, protein) or what upper-level biological ontologies should look like or do. It is not inconceivable that this process will take many years still.

1.9 Text mining and extraction

One common source of knowledge that many scientists wish to access is from the unstructured text of scientific publications. Due to the increasing volume of published articles, it is widely recognized (Hunter and Cohen, 2006) that researchers are unable to keep up with the flow of new research findings. Much expectation is placed on using computers to help researchers deal with this imbalance by mining the content for the relevant findings of each paper. However, there are many different things that can be mined out of research papers, and to do so completely and accurately is not possible today. Therefore, we will focus here only on the extraction of specific subsets of embedded information, including gene and biomolecule causal effects, molecular interactions and compartments, phenotype–gene associations, and disease treatments.

One way to mine content is simply to index key words and phrases based on text patterns and usage frequency. This is all search engines do, including Google. This does quite well in finding significant occurrences of words; however it fails to find exactly what is being said about something, that is, its semantics. For instance, indexing the phrase '... cytokine modulation may be the basis for the therapeutic effects of both anti-estrogens in experimental SLE.' One can readily identify *cytokine modulation* (CM) and its association with *therapeutic effects* (TE) or *experimental SLE* (xSLE), but the assertion that 'CM is a TE for xSLE' cannot be inferred from co-occurrence. Hence, limited knowledge about things being mentioned can be obtained using indexing, such as two concepts occurring in the same sentence, but the relation between them (if there is one) remains ambiguous.

Word-phrase indexing is very practical for search, but for scientific knowledge inquiries it is insufficient; what is specifically needed is the extraction of relations R(A, B). Although more challenging, there has been a significant effort invested to mine relations about specific kinds of entities from natural language. This is referred to as Information Extraction (IE), and it relies much more heavily on understanding some aspects of phrase semantics. Clearly this hinges on predefining classes of entities and sets of relations that are semantically mapped to each other (ontology). The objective of this is to quickly glean key relations about things like biomolecules, biostructures, and bioprocesses from research articles, thereby permitting the rapid creation of accessible knowledge bases (KBs) about such entities.

As an example, if one wanted to find out (from published research) if a particular gene is associated with any known disease mechanisms, one would

query the KB for any relation of the form ?gene ?rel [a Disease] (as well as [a Disease] ?rel ?gene). This form would allow any relation to a gene to be identified from the KB, where ?rel could mean 'is associated with,' 'influences,' 'suppresses,' or 'is over expressed in.' These relations should be defined in an ontology and the appropriate domain and range entity classes explicitly included. For IE to be most effective, it is useful to focus only on specific kinds of relations one is interested in, rather than trying to support a universal set. This helps reduce dealing with all the complexities of finding and interpreting specific relations from word-phrase patterns in natural languages, a problem that is far from being solved generically. Hence, it is desirable to have modules or cartridges for different IE target tasks, and which utilize different ontologies and controlled vocabularies.

Several open source or publicly accessible IE systems exist, including GATE, Geneways, OpenCalais, TextRunner, and OpenDMAP. OpenDMAP is specifically designed to extract predicates defined in the OBO system of ontologies (Relationship Ontology, RO), specifically those involved in protein transport, protein–protein interactions, and the cell-specific gene expression (Hunter *et al.*, 2008). They had applied it to over 4000 journals, where they extracted 72 460 transport, 265 795 interaction, and 176 153 expression statements, after accounting for errors (type 1 and type 2). Many of the errors are attributable to misidentification of genes and protein names. This issue will not be resolved by better semantic tools, since it is more basic and related to entity identification.

One possibility being considered by the community is that future publications may explicitly include formal identifiers for entities in the text, as well as controlled vocabularies, linked ontologies, and a specific predicate statement regarding the conclusion of the paper. Automated approaches that create such embedded assignments are being investigated throughout the research community, but so far show varying degrees of completeness and correctness, that is, both type 1 and type 2 errors. Much of this may be best avoided if the authors would include such embedding during the writing of their papers. Attractive as this sounds, it will require the development of easy to use and non-invasive tools for authors that do not impact on their writing practices. It will be quite interesting to follow the developments in this technology area over the next few years.

1.10 Gene expression

Gene Expression Analytics (GEA) is one of the most widely applied methodologies in bioinformatics, mixing data mining with knowledge discovery. Its advantage is that it combines experimentally controlled conditions with large-scale genomic measurements; as a technology platform has become commoditized so it can be applied cost effectively to large samples. Its weakness is that at best it is an average of many cells and cell types, which may be varied states, resulting in confounder effects; in addition, transcript levels usually do not

correspond to protein levels. It has become one of a set of tools to investigate and identify biomarkers that can be applied to the research and treatment of diseases. There is great expectation here to successfully apply knowledge-driven approaches around these applications, justifying the enormous investment in funded research and development to create knowledge to support the plethora of next-generation research.

GEA works with experimentally derived data, but shows best results when used in conjunction with gene annotations and sample-associated information; in essence, the expression patterns for many genomic regions (including multiple transcript regions per gene in order to handle splice variants) under various sample conditions (multiple affected individuals, genotypes, therapeutic perturbations, dosing, time-course, recovery). The data construct produced is an $N \times K$ matrix, $\boldsymbol{M}$, of expression levels, where N is the number of probes and K the number of samples. Much of the analytic enhancements have depended on performing appropriate statistics using replicate samples. This allowed the separation of variance arising from individual sample uncertainty and errors from the experimental factors that were being applied. The numeric data by itself can only support nearest neighbor comparisons, resulting in the construction of cluster trees: one for the relatedness of the expressing genes probed, the other for the relatedness of the tissue samples expressing the transcripts. Although many researchers try to find meaning in these cluster patterns (trees), they are for the most part artificial, due primarily to the experimental design, and have very little basis in normal biology.

More has been gained by associating additional knowledge to this matrix $\boldsymbol{M}$, such that both genes and samples have linked attributes that can be utilized in deeper analyses. Examples include utilizing the GO ontology for each gene to see if a correlation exists between the nearness of genes within the $\boldsymbol{M}$-derived gene cluster tree and the association of the genes with similar GO processes (Stoeckert *et al*., 2002). Other gene relations can be utilized as well common pathways (Slonim, 2002), common disease associations, and common tissue compartmentation. The same approach can be applied to the samples themselves, including variation in nutrients, genotypes (Cheung and Spielman, 2002), administered drugs, dosing, and clinical outcomes (Berger and Iyengar, 2009).

What is worthwhile remarking here is that all these different experimental design applications have a direct common correspondence with the mining of the numerically derived structures of the data in $\boldsymbol{M}$. That is, all the attributes associated with genes or samples can be viewed as formal relations linked to each instance or a gene or of a sample. For example, all genes G_i with attribute A_k can be evaluated for their possible correlation with higher expression values ($M_{ij} > z$) over all samples S_j by testing for $\mathrm{P}(M_{ij} > z | G_i.A_k)$, or even a subset of samples S_j with similar characteristics C_l, $\mathrm{P}(M_{ij} > z | G_i.A_k, S_j.C_l)$. As described earlier, this generalization supports many forms of knowledge mining, and therefore opens any applications of microarrays or biomarkers as fertile ground for KDD.

1.11 Pathways and mechanistic knowledge

Pathways are the abstraction of molecular mechanisms that are the basis for molecular functions and processes. They appear as graphs with directional flow; however the edges can have multiple meanings, such as catalyzed stoichiometric reactions (substrate, product as input, output respectively), protein signaling cascades and transcription factor and binding site activation/repression of genes. These structures can be further broken down into all the interactions each of the molecules participates in, yielding interaction graphs that are not obviously mapped to the process-oriented pathways. In each of these cases, the structures can be represented as graph objects, that is, sets of nodes connected by edges (Figure 1.1).

Fundamental to note is that pathways are not data; that is, they are not derived from single experiments. Rather, they are the result of hypothesis building and

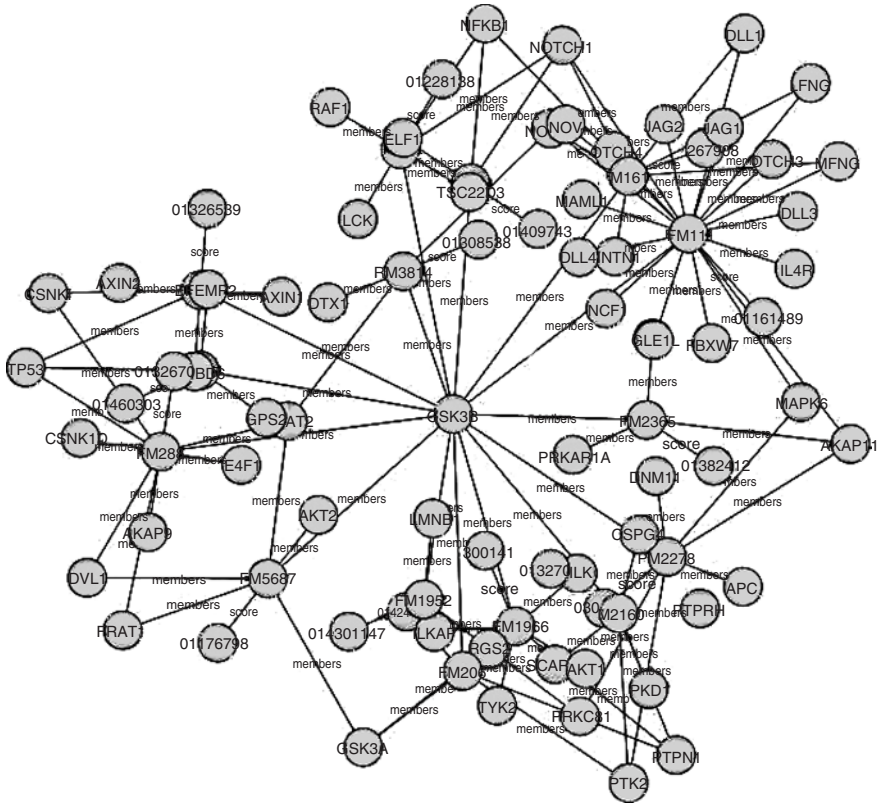

Figure 1.1 Assimilated knowledge around the GSK3b *gene from HEFalMp (Huttenhower et al., 2009), and its relations to other genes involved in similar processes.*

re-visioning over many experiments (and many years), and reflect our current knowledge of how biomolecules work together. In this regard, pathways are one manifestation of biological knowledge, and can be coded as a series of statements relating things to other things, that is, as graph objects with semantics (Losko and Heumann, 2009). The linked entities (nodes) can be mapped to existing protein and gene entities in other databases, thereby allowing pathways to be accessed and queried via entities from these other external sources (see Section 1.14.3).

Pathway structures lend themselves well to knowledge representations (*ADH1 catalyzes Ethanol_Acetylaldehyde_Reaction*), and can therefore aggregate additional facts and relations onto them. Annotations that describe possible regulations can be overlaid on top of these structures, possibly in another adjunct file (or graph). By using references to pathway models, it is possible to build layers of knowledge on top of more fundamental (canonical) pathways. These smart layers can include additional interactions, disease associations, and known polymorphic effects. Layers also provide a mechanism to describe and share additional knowledge that can be easily connected to existing pathway models via any software supporting the knowledge format. Reducing reliance on closed-vendor formats by increasing application independence is an important goal for making knowledge approaches successful.

The BioPax.owl ontology (currently as version 3) supports many kinds of relations within pathways (pathway steps, biochemical reactions, catalysis, modulation, part of complex, stoichiometric groups), and provides an exchange format supported by many pathway databases (Joshi-Tope *et al*., 2005). Basing pathway information on an ontology means it is semantically structured, not just syntactically. Not only can it be used to find substructures of interest, but one can also infer new relations from it based on existing relations. As an example, if a kinase K is known to down-regulate another signaling protein S, one can infer that all the downstream proteins of $S(D)$ will also be affected by the K kinase; this can be expressed as a rule and applied whenever one wishes to find all downstream affected components:

```
K down-regulates S AND S up-regulates D ⇒ K down-regulates D.
```

In addition, since molecular complexes are also modeled in BioPAX, a protein P which is part of a complex X, which is also part of another complex Y, is therefore also (P) part of complex Y, based on transitivity. Sets of these basic rules can be collected and applied strategically through a rule engine to help scientists find things such as potential disease mechanisms or candidate targets (Berger and Iyengar, 2009). Alternatively, mechanistic associations in conjunction with disease phenotypes can be used to explore novel drug applications (Qu *et al*., 2009).This form of knowledge discovery has only begun to be explored, and since its application potential is great, it will be important to see how it evolves in the coming years. Its success depends largely on how much material is accessible as digital knowledge, which recently has begun to look promising (see Section 1.14.3).

Often, pathways are used in parallel with GEA to better understand the roles and dynamics of various components under different conditions, including progression of cancer. By clustering genes by expression changes one can investigate if some pathways show correspondence of the encoded proteins with their location within a pathway structure, for instance downstream of a key control point. The summarization of many sources of evidence can be compiled into a comprehensive knowledge graph and used by researchers to find potential functionally associated gene relations (Huttenhower *et al*., 2009; see Figure 1.1). Pathway knowledge can also be combined with genotypic information to elucidate the effects gene variations have on mechanisms (Holford *et al*. 2009).

The tools of KDD can be applied here to help mine any possible correspondence between molecular interactions and classes of genes and their regulatory elements. Specifically, a pathway can be viewed as a proposed model structure S that can be the basis for analyzing expression dynamics using Bayesian network analysis $P(E|S,\theta)$. If one allows the model to evolve to better match expression data, the new relations can be interpreted as additional regulatory components on top of the pathway that were not known before.

1.12 Genotypes and phenotypes

As a final case study of knowledge-driven analysis, consider the complex problem of trying to identify relations between genotypes and phenotypes. This mapping has become increasingly relevant as researchers attempt to associate large numbers of human gene polymorphisms with observed traits, typically referred to as genome-wide associations (GWA). This analysis relies on the following existing resources in order to work: (1) detailed maps of genes in vicinities of polymorphisms, and (2) observed traits from familial studies that link sets of phenotypes and disorder symptoms to measured polymorphisms (Daly, 2010).

GWA by themselves cannot determine which specific gene is associated with a disorder, just that the influence seems to be nearby a polymorphism. Since there are multiple genes in these neighborhoods ($\sim$100 Mb), the actual affecting gene could be any one of these or even none of them. To further identify good gene candidates for the trait, one needs to analyze multiple sources of information and knowledge, including interactions and process membership (enrichment sets) with other proteins known to be associated with the related trait. Evidence may also come from animal models showing a related phenotype linked to the homologous gene. Then there might also be corroboration by way of common tissue expression for some of the genes, or even correlated expression changes due to a disease. In reality, there are many lines of reasoning with potential evidence, so the need for broad and flexible knowledge utilization is paramount.

Research into trait associations is ongoing and new evidence is continuously being generated. In order to validate an association, one must support an

accumulative knowledge model that can work with many kinds of relations and assertions, implying no hard defined data schema. This is beyond the capability of standard databases, and illustrates that traditional IT storage approaches will not be sufficient to support most forms of knowledge discovery. Once again this argues strongly for a linked relational approach such as using RDF.

Where one hopes this is heading is that one day we can bring together all forms of information and knowledge to any set of experimental data and hypotheses, and efficiently identify all possible logical explanations of a biological phenomenon, or the causes of a disease, or the possible treatments of a cancer, or the prediction of responses to a therapy based on an individual's genotype and lifestyle.

1.13 The Web's role in knowledge mining

We have seen how search has evolved over the last few years, and become a staple of general computer use – clearly the simpler interface is a critical factor in the case of Google. With the focus on knowledge, we should not lose sight of how the interface of a knowledge system will determine its utility for different groups of users. As a common example, the appearance of faceted browsing enables users to better assess overall content and quickly direct focus onto relevant subsections of it. This highlights a shift away from the 'magic one-query' paradigm to one of quickly zooming onto the critical subset of assertion.

Biomedicine is about combining many forms of information effectively towards the formation of predictive and causal models. The models themselves define relations that are based on our abstracted knowledge, which applies to many instances of systems: common molecular mechanisms in cells of similar tissue types; altered regulation of cell growth in normal and neoplastic cells; pathological similarities between human disease and animal models; variations in cellular processes arising from genotypic differences between individuals. Furthermore, many of our recent discoveries come from cross-pollination between areas: mathematical models of evolving networks and neural development; Green Fluorescent Protein (GFP) from algae and microscopy of cellular changes.

The scope of the required KD solutions must be broad enough to handle all the sources of information that biomedicine and life science researchers need in their increasingly interdisciplinary activities. That implies the utilization of ontologies that can bridge the concepts and information sets related to these sources. However, such ontologies should be defined not to impose our current snapshot of how we think biology works, but rather serve as a set of components, by which we can effectively describe new phenomena and derived hypotheses; in other words, ontologies for constructions of new proposed models and views.

Since knowledge should not be thought of as having imposed boundaries, the KD approaches offered must support the combination of data and ideas.

1.14 New frontiers

1.14.1 Requirements for linked knowledge discovery

As previously described, most information potentially available for knowledge applications is embedded within relational databases (RDBMSs). This is the most common queriable storage format, upon which enormous infrastructures, both public and private, have been built. RDBMSs have the following desirable benefits:

Information can be structured as needed, and defined in a schema.

The content is selected and cleaned before uploading to meet functional quality.

SQL is a well-defined and validated query algebra that can be optimized.

Access control to databases is managed.

They also have the following limitations that arise out of these same advantages:

Their applications are limited by the original design goals.

The content cannot be easily expanded to include new kinds of information.

The database schema typically has no means to apply formal logic.

The query capability is limited to within a single database.

It is not straightforward to combine multiple databases over a network.

These limitations have hindered advancing knowledge discovery in life science research environments by making the cost to extend systems prohibitively expensive[2]. More so, in a world where scientific data arises from many locations and with very different structures, scientists have been impeded from taking full advantage of available knowledge from the community via the Internet. Bioinformatics has a strong hacker component that is constantly writing widgets and shims to bridge existing but incompatible resources (Hull *et al*., 2004).

Several computer scientists have argued that ontologies are the solution to this problem, but as pointed to above, broad validated ontologies are not easy to create and do not trivially insert themselves into existing data systems. Ontologies represent a mechanism (along with a community process) for specifying content semantics, not a technology for bringing content together. The Linked Data model attempts to address most of these issues, providing a mapping to contemporary databases that supports federated queries and aggregation over the network. At the same time, linked data enable the incorporation of ontologies, even multiple ones. Two components necessary for data linking can be described using the concepts of data aggregation and data articulation, which now will be addressed.

1.14.2 Information aggregation

Aggregation is the process to efficiently find and correctly (semantically) link together information from multiple sources on the Web, based on relations

[2] Indeed this has spurred a culture of 'rebuild from scratch rather than build-upon.'

specified either in the content, or by a validated bridge source (look-up). This works best if the sources are already semantically defined and linking utilizes a referential linking model such as RDF triples. Aggregation also implies that not all information needs to be imported (i.e., massive data exchange) to build aggregate information, since the reference to a node is enough to state that all its other information is virtually connected as well. As an example, given the following two information sets:

```
PCKMz isa Kinase_Gene ; is_expressed_in Neural_Tissue  .

LTP is_necessary_for short-term_memory ; occurs_in
     Hippocampus_Tissue ;
is_caused_by local_Neural_Stimulation .
```

These two data sets (graphs) are disjoint (unconnected) except for the implied relation of `Hippocampus_Tissue` is a part of `Neural_Tissue`. If one now creates (or obtains from another researcher) the following statements:

```
PCKMz inhibited_by XC3751 . XC3751 blocks LTP . PCKMz
     has_role_in LTP .
```

Then once combined with the rest, these have the effect of connecting the previous statements (structurally and semantically) via the common `LTP` reference. Indeed, just knowing these facts is enough to connect the previous sets without having to explicitly transport any of the actual statements: their connections are available for discovery simply by accessing their URIs (e.g., all related information of LTP) through the Web.

I can formulate a query to ask: 'Which genes affect any kind of memory?' I could even do this using general concepts without using the specifics:

```
?aGene isa Gene; involved_in ?memProc . ?memProc isa
     Memory_Process .
```

Where the relation `involved_in` can be transitively inferred from the back-to-back `has_role_in` and `is_necessary_for` relations. The answer returned is:

```
PCKMz isa Gene .

PCKMz has_role_in LTP .

LTP is_necessary_for short-term_memory .

:= PCKMz involved_in short-term_memory .
```

These applications of aggregations do not happen trivially over the Web; reasoners require them to be pulled together in memory, either in one big batch, or incrementally as information is needed. Since there is no restriction (besides memory) on what one can aggregate, a researcher may wish to have all

necessary information locally for computational efficiency. Aggregation simply makes the task of finding, transporting and building linked data much more efficient without needing to develop complex software architectures *de novo*; this has been well illustrated by the Linked Open Data initiative.

1.14.3 The Linked Open Data initiative

As a specific example of data linking and aggregation on a grand scale, the Linked Open Data (LOD, http://linkeddata.org/) initiative, which began in Banff at the WWW2007 conference, is continuously aggregating many public data sources. At the time of this writing, LOD has grown to over 7.7 billion triples over 130 different data sources (http://esw.w3.org/topic/TaskForces/CommunityProjects/LinkingOpenData/DataSets/Statistics).

LOD addresses two points: (1) data from different sources are explicitly connected (rather than pages), and (2) the full set can be queried together based on their combined schema. LOD combines the notion of cloud computing and public access to connected and queriable data (not just Web pages). The URIs contained within serve as data identifiers that exactly pinpoint that record while at the same time are the points of linking between different data sets, for example ClinicalTrial and DrugBank. More recently, the Linked Open Drug Data (http://esw.w3.org/topic/HCLSIG/LODD) component of LOD just won the 2009 triplification award for largest and most useful data sets converted to RDF (http://triplify.org/Challenge/2009).

LOD is a live, Web-based proof-of-concept of what is possible if public data can be connected. Most of the original data relations are preserved, and new ones have been added between different sources when the relations where obvious. All these sets can now be queried using SPARQL (Figure 1.2, clinical trials data at http://linkedCT.org rendered with the Cytoscape S*QL plug-in). LOD is to serve as a starting point as to what is possible within the public Web. As discussed earlier, large-scale knowledge discovery requires robust and structured accessibility to large sets of information, complete with semantic relations and associated class definitions. LOD is helping point out how to move beyond our current reliance on local databases, and prepare us to begin using structured information on the Web to help solve complex, real-world challenges.

1.14.4 Information articulation

Pulling together information from different sources only gets you so far; information can be collected, but it may be in a form that does not provide significant value to specific inquiries, such as new insights that could arise from combined relations between instances. This is a limitation with the current forms of linked data, where data is available as a collected set, but does not necessarily offer any deeper logical insights. This limitation is primarily semantic, and indicates that there is more to knowledge mining than information models; additional logical relations are usually required to pose deeper inquiries.

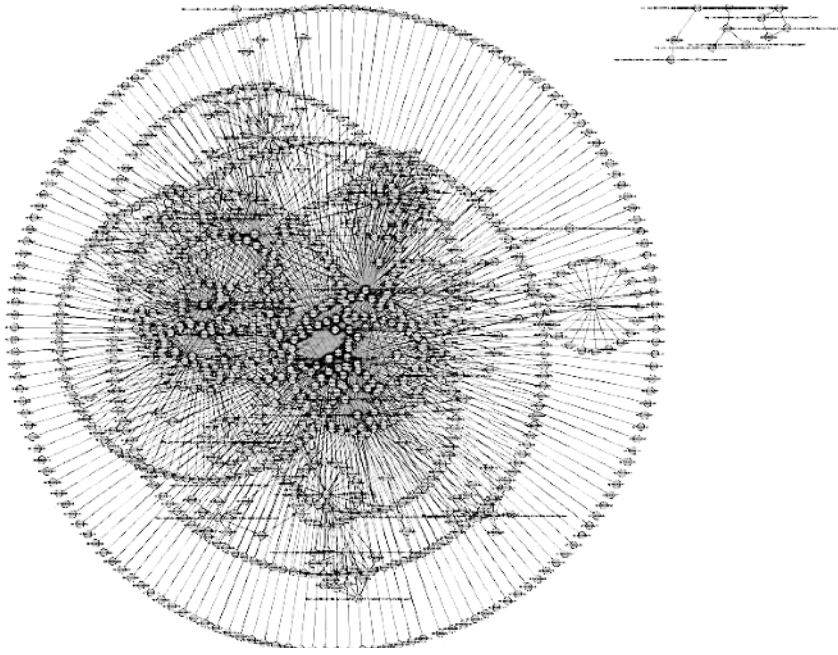

Figure 1.2 Results of a SPARQL query on LinkedCT, a LOD resource from ClinicalTrials.gov, showing all the cancer clinical trials surrounding their respective specific cancer groups. Concentricity indicates studies investigating multiple cancers together.

For example, if a data set of customers can be linked to a data set of products from multiple vendors, one can readily query and find which customers tend to buy specific things from certain vendors, and which vendors offer products sold most frequently. However the insight into which vendor products are sold together (product–product relation) requires additional analytics and an expression to represent such novel associations (e.g., bought_with). The ability to infer new meaningful relations from existing semantics and incorporate these as additional assertions into the existing knowledge set is defined here as *information articulation*. The new relations increase the degrees of freedom of data nodes and thereby the reusability of data sets; in essence, the new relations *articulate the data*.

Information articulation can be best compared to the use of a spreadsheet whereby a user inserts a new column whose cell values are the direct result (computational output) of other existing columns from the same or different worksheets. However, rather than cells being created with new values, existing nodes (entities) are being linked by new (inferred) relations, or new nodes are being created to hold N-ary (fused) relations. This can be in the form of transitive closure (if A is part of B and B is part of C, then A is part of C) or other

common rules (if Gene G is associated with Type 2 Diabetes, and Gene G is part of Pathway P, then Pathway P is involved in Type 2 Diabetes). The articulation takes the form of a rule: sets of specific relations create new deeper relations. The knowledge of when to apply the rule comes from a deeper domain understanding. Such knowledge may be defined within ontologies, but often a complex rule set cannot be coded using standard ontological frameworks.

Articulation is usually monotonic, only adding new relations to existing ones, without any deletions. The goal of information articulation is to create new usable insights by making novel connections between existing knowledge and new empirical data. Since novel relations must always be built from/on top of existing substrates, there needs to be flexibility for adding them onto existing structures, and this is what enables articulation. Another way to think about it is that when one performs inferences on knowledge, the resulting 'new findings' need to be connected with the input information so that they are retrievable and remain in context. These relations must make semantic sense for their respective domains.

Information articulation requires a sufficient degree of domain semantics to be known in order to make such inferences from initial entity sources (son-father-brother $\rightarrow$ uncle). This also implies that one cannot know everything about the information model in advance; the extension of knowledge needs to be done in such a way that it does not create inconsistencies (UNCLE $\not\subset$ WOMAN). Information articulation also helps differentiate between *extensional sets* (data facts) and *intensional sets* (implied by semantics and reasoning). Intensional sets are often more endowed with usable meaning for future inferencing. Knowledge discovery will require that we can build upon facts by using tools and structures that can further articulate information with usable meaning.

1.14.5 Next-generation knowledge discovery

We have presented an overview of the relation between information and knowledge and how one can computationally apply knowledge to information to create new knowledge. Much of the theory has emerged from decades of research, but in the last several years, the combined expansion of the Web in conjunction with advances in semantic standards and technologies have dramatically opened the possibility of performing knowledge discovery on a global scale. Bioinformatics is one domain that can begin to take advantage of these resources to advance our understanding of biology and help us conquer diseases.

The key elements of knowledge discovery – logic, abstraction, relational graphs, and statistical induction – must be combined in consistent and flexible ways in order to address the richness of biomedical information. Once all of the biomedical information has been structurally and semantically normalized, many kinds of applications become possible. Comprehension of biological mechanisms will be critical for new therapeutics, and knowledge of one disease will often be key to the understanding of a very different disease. Biology itself is highly interlinked, so we will require information tools and logic that can match these structures.

In the near future, we will see new knowledge coming from multiple communities over many projects. Standards for information are important, but simply agreeing on the same formats and ontologies will not be sufficient. Applications that search, aggregate, and analyze diverse information will have to apply deeper logic fundamentals that cannot be addressed by ontologies alone. Information articulation will be key to performing deeper inductive-driven discovery across large volumes of information.

Pushing these new paradigms in the right direction is essential. New standards need to used and assessed, and if validated, the community must be encouraged to incorporate them into all our various resources, public as well as local. We still have a long way to go for increased awareness and general acceptance. For the most part, this will be driven by successful demonstrations of these information technologies in the context of scientific research and as part of the scientific process. In other cases they will become incorporated in commercial technologies driven by new opportunities in biotechnology research and personalized medicine. However, in all cases they will be chosen to realize benefits and improvements to knowledge discovery. A scientific community based on the sharing and commerce of knowledge discovery practices is closer at hand now than at any previous time in our history.

1.15 References

Berger, S.I. and Iyengar, R. (2009) Network analyses in systems pharmacology. *Bioinformatics*, **25**(19), 2466–72.

Cheung, V.G. and Spielman, R.S. (2002) The genetics of variation in gene expression. *Nat. Genet.*, **32**, 522–5.

Ciccarese, P., Wu, E., Wong, G., *et al*. (2008) The SWAN biomedical discourse ontology. *J Biomed Inform.*, **41**(5), 739–51.

Daly, A.K. (2010) Genome-wide association studies in pharmacogenomics. *Nat Rev Genet.*, **11**(4), 241–6.

Fagin, R., Halpern, J.Y., Moses, Y., and Vardi, M.Y. (1995) *Reasoning About Knowledge*, MIT Press.

Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R. (eds) (1996) *Advances in Knowledge Discovery and Data Mining*, AAAI Press.

Getoor, L., Friedman, N., Koller, D., *et al*. (2007) Probabilistic relational models, in *Introduction to Statistical Relational Learning* (eds. L. Getoor and B. Taskar), MIT Press, pp. 129–74.

Holford, M.E., Rajeevan, H., Zhao, H., *et al*. (2009) Semantic web-based integration of cancer pathways and allele frequency data. *Cancer Inform.*, **8**, 19–30.

Horrocks, I., Kutz, O., and Sattler, U. (2006) The even more irresistible SROIQ, in *Proceedings of the 10th International Conference on Principles of Knowledge Representation and Reasoning (KR2006)*, AAAI Press, pp. 57–67.

Hull, D., Stevens, R., Lord, P., and Goble, C. (2004) Integrating bioinformatics resources using shims. Twelfth International Conference on Intelligent Systems for Molecular Biology (ISMB2004), Glasgow, UK.

Hunter, L. and Cohen, K.B. (2006) Biomedical language processing: what's beyond PubMed? *Mol Cell*, **21**, 589–94.

Hunter, L., Lu, Z., Firby, J., *et al*. (2008) OpenDMAP: an open source, ontology-driven concept analysis engine, with applications to capturing knowledge regarding protein transport, protein interactions and cell-type-specific gene expression. *BMC Bioinformatics*, **9**, 78.

Huttenhower, C., Haley, E.M., Hibbs, M.A., *et al*. (2009) *Exploring the Human Genome with Functional Maps*, Cold Spring Harbor Laboratory Press.

Joshi-Tope, G., Gillespie, M., Vastrik, I., *et al*. (2005) Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res.*, **33**(Database issue), D428–32.

Kessler, N. Schuhmann, H., and Morneweg, S. (2004) The linear pentadecapeptide gramicidin is assembled by four multimodular nonribosomal peptide synthetases that comprise 16 modules with 56 catalytic domains. *J. Biol. Chem.*, **279**(9), 7413–19.

Losko, S. and Heumann, K. (2009) Semantic data integration and knowledge management to represent biological network associations. *Methods Mol. Biol.*, **563**, 241–58.

Naisbitt, J. (1982) *Megatrends*, Avon Books.

Nature Editorial (2009) Data's shameful neglect. *Nature*, **461**, 145.

Neumann, E.K. and Prusak, L. (2007) Knowledge networks in the age of the Semantic Web. *Brief. Bioinformatics*, **8**(3), 141–9.

Prusak, L. and Davenport, T. (2000) *Working Knowledge: How Organizations Manage What they Know*, Harvard University Business School Press.

Qu, X.A., Gudivada, R.C., Jegga, A.G., *et al*. (2009) Inferring novel disease indications for known drugs by semantically linking drug action and disease mechanism relationships. *BMC Bioinformatics*, **10**(Suppl 5), S4.

Rubin, D.L., Lewis, S.E., Mungall, C.J., *et al*. (2006) National Center for Biomedical Ontology: advancing biomedicine through structured organization of scientific knowledge. *OMICS* **10**(2), 185–98.

Slonim, D.K. (2002) From patterns to pathways: gene expression data analysis comes of age, *Nat. Genet.*, **32**, 502–508.

Stoeckert, C.J., Causton, H.C., and Ball, C.A. (2002) Microarray databases: standards and ontologies. *Nat. Genet.*, **32**, 469–73.