Annotating Protein Function through Lexical Analysis

Rajesh Nair and Burkhard Rost

■ We now know the full genomes of more than 60 organisms. The experimental characterization of the newly sequenced proteins is deemed to lack behind this explosion of naked sequences (sequence-function gap). The rate at which expert annotators add the experimental information into more or less controlled vocabularies of databases snails along at an even slower pace. Most methods that annotate protein function exploit sequence similarity by transferring experimental information for homologues. A crucial development aiding such transfer is large-scale, work- and management-intensive projects aimed at developing a comprehensive ontology for gene-protein function, such as the Gene Ontology project. In parallel, fully automatic or semiautomatic methods have successfully begun to mine the existing data through lexical analysis. Some of these tools target parsing controlled vocabulary from databases; others venture at mining free texts from MEDLINE abstracts or full scientific papers. Automated text analysis has become a rapidly expanding discipline in bioinformatics. A few of these tools have already been embedded in research projects.

Proteins are the machinery of life. The information for life is stored in a four-letter alphabet in the genes (deoxyribonucleic acid [DNA]) (Alberts et al. 1994; Lodish et al. 2000). This four-letter DNA alphabet is translated into a 20-letter amino acid alphabet constituting the basic language for proteins, the machinery of life. Proteins are formed by joining amino acids through peptide bonds. Proteins differ greatly in the number of amino acids joined (from 30 to more than 30,000) and the arrangement and types of amino acids used (dubbed residues when joined in proteins). *Proteins* are the macromolecules that perform all important tasks in organisms, such as catalysis of biochemical reactions, transport of nutrients, and recognition and transmission of signals. The plethora of role aspects of any particular protein is referred to as its *function*. However, protein function is not a well-defined term. Instead, function is a complex phenomenon that is associated with many mutually overlapping levels: chemical, biochemical, cellular, organism mediated, developmental, and physiological. These levels are related in complex ways; for example, protein kinases can be related to different cellular functions (such as cell cycle) and to a chemical function (transferase) plus a complex control mechanism by interaction with other proteins. The same kinase can also be the culprit that leads to misfunction, or disease. Thus, identifying protein function is a step toward understanding diseases and identifying drug targets (Brutlag 1998).

The first entire genome (DNA) sequence of a free living organism, *Haemophilus influenzae,* was published in 1995 (Fleischmann et al. 1995). Currently, we know the full genomes for more than 100 organisms; for more than 60 of these, the data are publicly available and contribute about 250,000 protein sequences, that is, about one-fourth of all currently known protein sequences (Carter et al. 2003; Liu and Rost 2001; Pruess et al. 2003). The number of entirely sequenced genomes is expected to continue growing exponentially for at least the next few years. This explosion of se-

```
ID   MYOD_HUMAN     STANDARD;      PRT;   319 AA.
AC   P15172;
. . .
DE   Myoblast determination protein 1 (Myogenic factor MYF-3).
GN   MYOD1 OR MYF3.
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
OX   NCBI_TaxID=9606;
. . .
CC   -!- FUNCTION: INVOLVED IN MUSCLE DIFFERENTIATION (MYOGENIC FACTOR).
CC       INDUCES FIBROBLASTS TO DIFFERENTIATE INTO MYOBLASTS. ACTIVATES
CC       MUSCLE-SPECIFIC PROMOTERS. INTERACTS WITH AND IS INHIBITED BY THE
CC       TWIST PROTEIN. THIS INTERACTION PROBABLY INVOLVES THE BASIC
CC       DOMAINS OF BOTH PROTEINS (BY SIMILARITY).
CC   -!- SUBUNIT: EFFICIENT DNA BINDING REQUIRES DIMERIZATION WITH ANOTHER
CC       BHLH PROTEIN. SEEMS TO FORM ACTIVE HETERODIMERS WITH ITF-2.
CC   -!- SUBCELLULAR LOCATION: Nuclear.
CC   -!- SIMILARITY: BELONGS TO THE BASIC HELIX-LOOP-HELIX (BHLH) FAMILY OF
CC       TRANSCRIPTION FACTORS. “MYOGENIC FACTORS” SUBFAMILY.
. . .
DR   SWISS-2DPAGE; GET REGION ON 2D PAGE.
KW   Myogenesis; Differentiation; Developmental protein; Nuclear protein;
KW   Transcription regulation; DNA-binding.
FT   DNA_BIND    109    121       BASIC DOMAIN.
FT   DOMAIN      122    161       HELIX-LOOP-HELIX MOTIF (BY SIMILARITY).
FT   CONFLICT    124    124       K -> E (IN REF. 2).
. . .
```

Figure 1. Protein Entry in SWISS-PROT.

The SWISS-PROT identifier for the protein MYOD_HUMAN is found under the header ID. The type of protein and its source organism are found under the DE and OS headers, respectively. Detailed functional information regarding the protein is found under the header CC. This information is written in plain English and is not suitable for computer analysis. Following the KW header are keywords describing the function of the protein. The keywords use a restricted vocabulary and are ideal for tools for text analysis.

quence information has widened the gap between the number of protein sequences deposited in public databases and the experimental characterization of the corresponding proteins (Baker and Brass 1998; Koonin 2000; Lewis et al. 2000; Rost and Sander 1996). Bioinformatics, sometimes referred to as *functional genomics,* plays a central role in bridging the sequence-function gap through the development of tools for faster and more effective prediction of protein function (Bork et al. 1998; Fleischmann et al. 1999; Luscombe et al. 2001; Valencia 2002; Valencia and Pazos 2002).

Here, we briefly review some of the attempts at annotating function through homology transfer. The most widely used methods that allow guessing protein function rely on the ability to correctly mine the information deposited in public databases and scientific journals. Although we need a comprehensive ontology for protein function, developers have begun to successfully explore the marvels of an ever-increasing body of research in biology and medicine. There are two major methods attempting automatic lexical analysis: (1) parsing of controlled vocabulary from databases and (2) mining unstructured text available from scientific publications. We could not cover all the promising approaches that have mushroomed over the last 5 to 10 years; therefore, we focus in detail on a few success stories.

Annotations and Annotation Transfer of Protein Function

Information about protein sequences is stored in public databases such as SWISS-PROT and TrEMBL (table 1). SWISS-PROT (Apweiler 2001; Bairoch and Apweiler 2000) is an expert-curated database that also contains annotations about function (figure 1). These annotations

Database	URL
SWISS-PROT (Boeckmann et al. 2003)	www.ebi.ac.uk/swissprot
TrEMBL (Boeckmann et al. 2003)	www.ebi.ac.uk/trembl
Gene Ontology (Ashburner et al. 2000)	www.geneontology.org
MIPS (Mewes et al. 2000)	mips.gsf.de
PEP (Carter et al. 2003)	cubic.bioc.columbia.edu/db/PE1

Table 1. Web Sites of Major Databases and Genome Resources.

are added by a team of expert annotators who extract this information primarily from journal publications (Junker et al. 2000). TrEMBL (Bairoch and Apweiler 2000) consists of entries that are derived from the translation of all coding sequences in the EMBL nucleotide sequence database that are not in SWISS-PROT. Unlike SWISS-PROT records, those in TrEMBL are awaiting manual annotation. SWISS-PROT currently contains only 113,470 sequence entries, and the TrEMBL database contains over 755,169 sequence entries (Boeckmann et al. 2003).

Annotations of function primarily occur through homology transfer. Experimentally determining protein function continues to be a laborious task that can take enormous resources; for example, more than a decade after the discovery, we still do not know the precise and entire functional role of the prion protein (Harrison et al. 1997). The automatic elucidation of the protein function is therefore an appealing challenge (Apweiler et al. 1997; Eisenberg et al. 2000; Gaasterland and Sensen 1996). The most commonly used approach for the automatic elucidation of protein function relies on the fact that two proteins with similar sequence often have a similar function.

The basic idea to exploit this fact involves the following steps: (1) extract the experimental information from the literature into a controlled vocabulary of annotated databases; (2) establish thresholds T for pairwise sequence similarity that imply similarity in function; (3) for a protein U of unknown function, search the database for proteins $\{K\}$ that have a sequence similarity to U: $SIM(K, U) > T$; and (4) if any such protein K is found, transfer its annotation to U. Albeit this concept appears straightforward, in practice, there are many hurdles to overcome: First, the function is not well defined; hence, it is very difficult to create controlled vocabularies (Ashburner et al. 2000; Bairoch and Apweiler 2000). Second, because function is such a complex phenomenon, it is very difficult to assign one number that describes all these roles at once (Ashburner et al. 2000; Todd et al. 2001). Third, to add to the complication, it seems that the precise values for thresholds of significant sequence similarity (T) are actually specific to particular function—that is, become $T(F)$—and have to be reestablished for any given task (Devos and Valencia 2000; Nair and Rost 2002a, 2002b; Ouzounis et al. 1998; Rost 2002, 1999; Todd et al. 2001; Wilson et al. 2000; Wrzeszczynski and Rost 2003). In general, the inference of function is reliable only for very high levels of sequence similarity (Devos and Valencia 2001; Nair and Rost 2002; Rost 2002). For reliably inferring the subcellular localization of a protein using homology transfer, a sequence identity of more than 80 percent is required. At this sequence identity, subcellular localization annotations can be transferred at more than 90-percent accuracy. Below this threshold, the accuracy of annotation transfer rapidly decreases (Nair and Rost 2002b).

Several pitfalls in transferring annotations of function have been reported, for example, having inadequate knowledge of thresholds for "significant sequence similarity"; using only the best database hit; or ignoring the domain organization of proteins (Bork and Koonin 1998; Devos and Valencia 2001; Doerks et al. 1998; Galperin and Koonin 2000). Despite all these problems, the majority of annotations about function in public databases result from homology transfer (Devos and Valencia 2001; Koonin 2000; Valencia 2002). Databases such as SWISS-PROT usually do not provide pointers for the origin of the information. One problem arising from this approach is that it is difficult to distinguish annotations that are experimental from those that are predicted.

Problem 1: Multiple Levels of Description

The function of a protein is context dependent. Database annotations of protein function are often confusing because of the variety of functional roles (Attwood 2000). For reliable automatic predictions, computer-readable hierarchical descriptions of function are needed (Bork et al. 1998; Overbeek et al. 1997).

Several groups and associations have ventured to introduce numeric schemata to define function. The first attempt was the introduction of enzyme classification numbers (Webb 1992); this classification uses four digits to classify enzymatic activity. The first enzyme classification digit distinguishes the general types of enzymes; the second enzyme classification digit specifies the substrate (oxireductases), the group transferred (transferases), the type of bond (hydrolases, lyases, ligases), or the type of reorganization (isomerases). The third and fourth digits provide more detail (for an excellent survey of structural aspects of enzymatic function, see Todd, Orengo and Thornton [2001]). MIPS attempts to extend this idea to a wider perspective, with more proteins and more roles, through its classification catalog (Mewes et al. 2000).

Arguably, the most impressive gargantuan effort at defining ontology for protein function originates from the gene ontology consortium (Ashburner et al. 2000). Gene ontology distinguishes three levels of protein function: (1) molecular, (2) biological, and (3) cellular. At the molecular level, the protein can, for example, catalyze a metabolic reaction and recognize or transmit a signal. In a biological process, a set of many cooperating proteins is responsible for achieving broad biological goals, for example, mitosis, purine metabolism, or signal transduction cascades. The cellular category includes the structure of subcellular compartments, the localization of proteins, and macromolecular complexes. Examples include nucleus, telomere, and origin recognition complex. The subcellular localization of a protein is an essential attribute for this level. The totality of the physiological subsystems and their interplay with various environmental stimuli determine properties of the phenotype, the morphology and physiology of the organism and its behavior. Gene ontology is not complete. In fact, now after almost a decade of efficient work, the first notable coverage of the experimental information is complete, and the developers contemplate restarting. Nevertheless, gene ontology constitutes the best set of definitions available today.

Problem 2: No Machine-Readable Functional Information

Nearly all databases present the protein sequence in formats that are more or less straightforward for parsing by computers. However, annotations are mostly written in plain text using a rich biological vocabulary that often varies in different areas of research (figure 1). Such annotations are primarily meant for the eyes of human experts; hence, they are not machine readable (Eisenhaber and Bork 1999). Another problem that hampers automatic annotations is the quality of database annotations: Only a few database groups attempt a quality control of curated annotations (Tsoka and Ouzounis 2000).

Automatic Lexical Analysis of Controlled Vocabularies

Protein databases such as SWISS-PROT usually contain functional annotations at a very detailed level of biochemical function; for example, a given sequence is annotated as a *cdc*2 kinase but not as being involved in intracellular communication (Apweiler 2001; Tamames et al. 1998). A number of text-analysis tools have been implemented that infer various aspects of cellular function from database annotations of molecular function. Many methods explore the functional annotations in SWISS-PROT, especially the keyword annotations (Eisenhaber and Bork 1999; Fleischmann et al. 1999; Nair and Rost 2002a; Tamames et al. 1998).

SWISS-PROT currently contains over 800 keyword functional descriptors. Semantic analysis of the keywords is used to categorize proteins into classes of cellular function (Andrade et al. 1999b; Bork et al. 1992; Karp et al. 1999; Ouzounis et al. 1996; Riley 1993; Riley and Labedan 1997). There are two types of methods: (1) fully automated and (2) semiautomatic.

With fully automated methods, the problem of automatically extracting rules from keywords has parallels to the problem of *text categorization*, that is, assigning predefined categories to free-text documents. Many statistical learning methods have been applied to this problem, including nearest-neighbor classifiers (Yang and Pederson 1997), multivariate regression models (Schutze et al. 1995; Yang and Chute 1992), probabilistic Bayesian models (Lewis and Ringuette 1994), symbolic rule learning (Apte et al. 1994), and *m*-ary (multiple-category) classifiers such as the *k*–nearest neighbor (Dasarathy 1991) and the linear least squares fit (LLSF). These methods have been intensively studied and are among the most accurate for text categorization (Yang and Liu 1999). The majority of the tools for annotating function are based on one of these methods.

Database	URL
LOCkey (Nair and Rost 2002)	cubic.bioc.columbia.edu/services/LOCke
GeneQuiz (Tamames et al. 1998)	jura.ebi.ac.uk:8765/ext-genequiz/
Meta_A (Eisenhaber and Bork 1999)	mendel.imp.univie.ac.at/CELL_LC
AbXtract (Andrade and Valencia 1998)	columba.ebi.ac.uk:8765/andrade/a
SUISEKI (Blaschke and Valencia 2002)	www.pdg.enb.uam.es/suiseki/

Table 2. Resources for Text Analysis.

Some of the major methods for annotating function are LOCKEY (Nair and Rost 2002), SPEARMINT (Bazzan et al. 2002; Kretschmann et al. 2001), and the support vector machine (SVM)-based approach of Stapley et al. (2002) (table 2).

Semiautomated methods are based on building dictionaries of rules. Keywords characteristic of each of the functional classes are first extracted from a set of classified example proteins. With these keywords, a library of rules is created that associates a certain pattern of occurrence of keywords to a functional class. The major methods in this category are EUCLID (Tamames et al. 1998), META_A (Eisenhaber and Bork 1999), and RULEBASE (Fleischmann et al. 1999) (table 2). We review the LOCKEY and EUCLID algorithms as examples of the two main approaches.

The LOCKEY system (Nair and Rost 2002a) is a novel *m*-ary classifier that predicts the subcellular localization of a protein based on SWISS-PROT keywords. The algorithm can be divided into two steps (figure 2): (1) building data sets of trusted vectors for known proteins and (2) classifying unknown proteins. First, a list of keywords is extracted from SWISS-PROT for all proteins with known subcellular localization. On average, most proteins have between two and five keywords. A data set of binary vectors (Salton 1989) is generated for each protein by representing the presence of a certain keyword in the protein by 1 and absence by 0. Second, to infer subcellular localization of an unknown protein *U,* all keywords for *U* are read from SWISS-PROT. These keywords are translated into a binary keyword vector. From this original keyword vector, LOCKEY generates a set of all possible combinations of alternative vectors by flipping vector components of value 1 (presence of keyword) to 0 in all possible combinations. For example, for a protein with three keywords, there are $2^3 - 1 = 7$ possible subvectors: 111, 110, 101, 011, 100, 010, and 001. These subvectors constitute all possible keyword combinations for protein *U*. The keyword combination, that is, subvector, that yields the best classification of *U* into 1 of 10 classes of sub-cellular localizations is found. All exact matches of each of the subvectors to any of the proteins in the trusted set are retrieved by finding all proteins in the trusted set that contain all the keywords present in the subvector. By construction, the proteins retrieved in this way can also contain keywords not found in *U*.

The next task is to estimate the surprise value of the given assignment. Toward this end, LOCKEY simply compiles the number of proteins belonging to each type of subcellular localization. This procedure is repeated in turn for each of the subvectors, and localization is finally assigned to a protein by minimizing an entropy-based objective function. The system accurately solves the classification problem when the number of data points (proteins) and the dimensionality of the feature space (number of keywords) are not too large. LOCKEY reached a level of more than 82-percent accuracy in a full cross-validation test. However, because of a lack of functional annotations, the system failed to infer localization for more than half of all proteins in the test set (**Note:** A number of SWISS-PROT keywords are biologically correlated to subcellular localization; for example, DNA-binding proteins always localize to the nucleus. These keywords, which were all simple one-to-one correlations, were excluded from testing because the goal was to estimate the true ability of the algorithm to infer complex correlations among the keywords). For five entirely sequenced proteomes—(1) *Saccharomyces cerevisiae* (yeast), (2) *Caenorhabditis elegans* (worm), (3) *Drosophila melanogaster* (fly), (4) *Arabidopsis thaliana* (plant), and (5) a subset of all human proteins—the LOCKEY system automatically found about 8000 new annota-

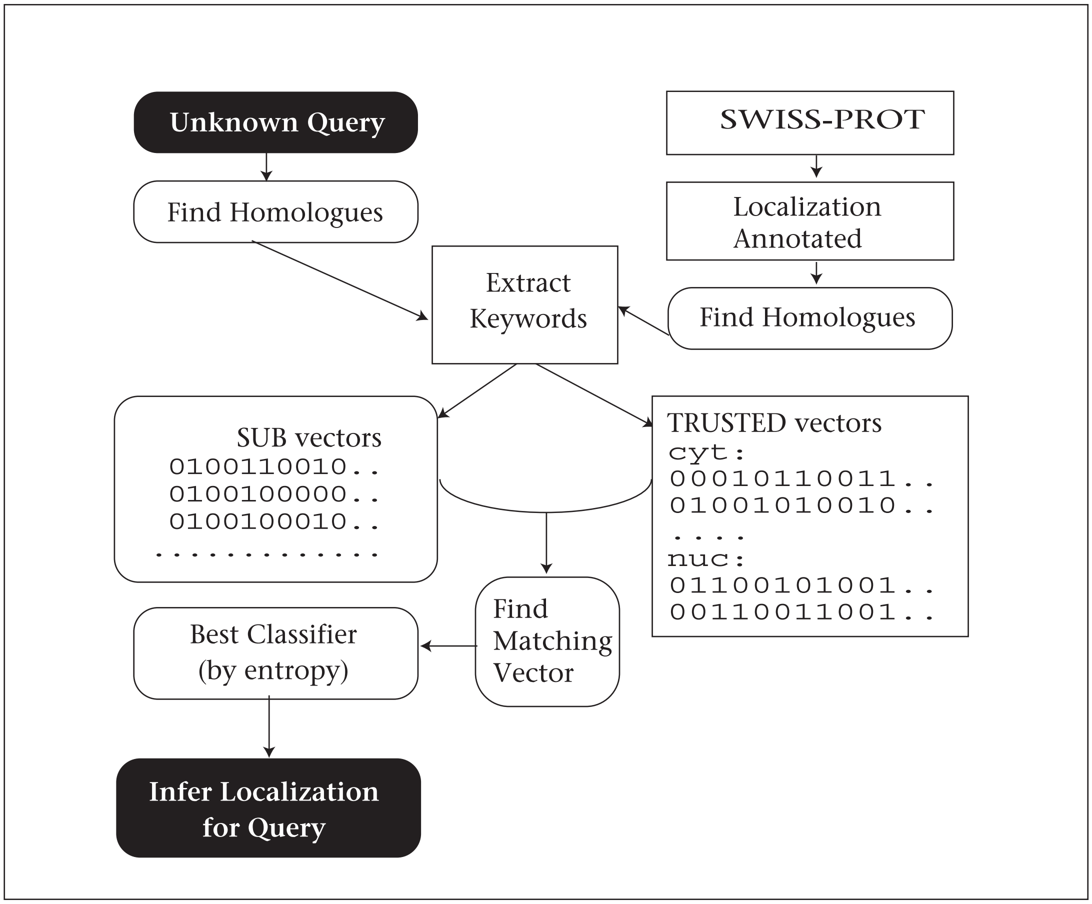


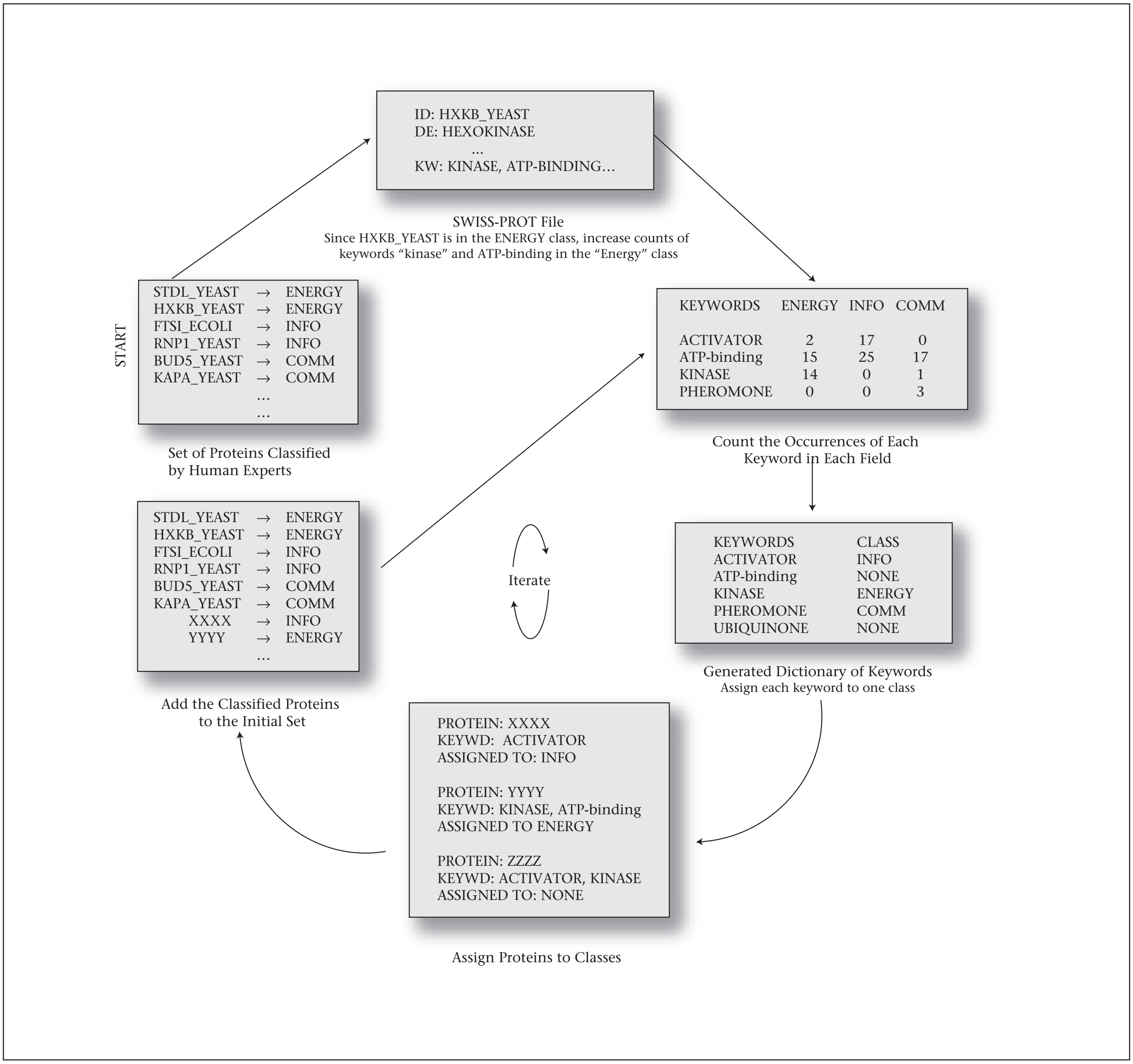
Figure 2. The LOCKEY Algorithm.

A sequence-unique data set of localization-annotated SWISS-PROT proteins was first compiled. Keywords were extracted for these proteins and merged with any keywords found in homologues. The keywords were represented as binary vectors in the trusted vector set. An unknown query was first annotated with keywords through identification of SWISS-PROT homologues. Keywords for the query were represented as binary vectors. All possible keyword combinations were constructed (the subvectors). The best-matching vector was found based on entropy criteria (Nair and Rost 2002a). This vector was used to infer localization for the query.

tions about subcellular localization. LOCKEY has been optimized to provide fast annotations. Annotating the entire *C. elegans* proteome took less than 4 hours on a Pentium III 900-megahertz machine. The algorithm is limited to problems with relatively few data points (proteins) in the vector set (*n* << 1000000) and with few keywords (*n* << 10000).

The EUCLID system (Tamames et al. 1998) uses SWISS-PROT keywords to classify proteins into 14 classes of cellular function according to the scheme originally proposed by Monika Riley (Karp et al. 1999; Krawiec and Riley 1990; Riley 1993; Riley and Labedan 1997). The 14 classes are grouped into 3 broad functional classes: (1) energy, (2) information, and (3) communication. The EUCLID system can be summarized as followed: First, keywords characteristic of each of the functional classes are extracted from a set of classified example proteins provided by a human expert. This dictionary of characteristic keywords satisfies the following criteria: (1) only keywords with functional meaning are used, and keywords with no functional information are excluded (for example, hypothetical or three-dimensional structure); (2) only keywords appearing in more than one SWISS-PROT entry are considered; and (3) only keywords with more than 85 percent of their occurrences in a single functional category are included in the dictionaries.

Next, to assign sequences to classes, a simple voting scheme is used. A sequence is automatically classified in the functional class to which the majority of its keywords belong. The dictionary of keywords is then used to automatically assign all proteins from the database if a sufficient match is found. Proteins thus assigned to a functional class are analyzed to extract a new, more extensive dictionary of characteristic key-



KEYWORDS	ENERGY	INFO	COMM
ACTIVATOR	2	17	0
ATP-binding	15	25	17
KINASE	14	0	1
PHEROMONE	0	0	3

KEYWORDS	CLASS
ACTIVATOR	INFO
ATP-binding	NONE
KINASE	ENERGY
PHEROMONE	COMM
UBIQUINONE	NONE

Figure 3. The EUCLID *Algorithm.*

Scheme of the iterative method used to classify sequences in three functional classes. The classification relies on the definition of a dictionary of keywords characteristic for a particular functional class. First, experts assign hexokinase from yeast (hxkb_yeast) to the energy class. Second, a keyword dictionary is constructed scoring the keywords associated with hexokinase in the energy class. Third, the same dictionary is then extended to classifying other proteins. The process is iterated until no more keywords are gained.

words. The process is iterated until the classification quality no longer increases (figure 3). A limitation of this approach is that only simple correlations between keywords can be discovered. The method is easily scalable and can be applied to very large protein databases. For the genome sequence of *Mycoplasma genitalium* (Fraser et al. 1995), the EUCLID system was able to classify 52 percent of the sequences at a classification accuracy of 82 percent. The EUCLID algorithm has been incorporated into the GENEQUIZ workbench for sequence analysis (Andrade et al. 1999a). GENEQUIZ is a semiautomated protein sequence analysis workbench whose principal purpose is to infer a specific and reliable functional assignment together with a broad cellular role for a query protein by analyzing annotations from sequence database matches.

Mining Free Text from the Literature

Experimental results are usually published first in scientific journals. Because such publications do not conform to any standardized rules, this information is not computer readable. At best, this lack of automation leads to a severe delay in incorporating the information into databases. Furthermore, a lot of the data will lie buried in the literature forever. One solution to this problem will be to adopt standards similar to the PDB model for protein structure for depositing functional information into databases, that is, requiring deposition of, say, protein-protein interactions into a public database along with publication.

However, currently, mining text is the only way of retrieving functional information from the literature. In recent years, many groups have worked on dedicated problems in this area, such as machine selection of articles of interest (Iliopoulos et al. 2001; Shatkay et al. 2000), automated extraction of information using statistical methods (Stapley and Benoit 2000; Stephens et al. 2001), and natural language–processing techniques (Friedman et al. 2001; Ng and Wong 1999; Thomas et al. 2000; Yakushiji et al. 2001) as well as setting up specialized knowledge bases for storing molecular knowledge (Stevens et al. 2000). The invaluable electronic availability of scientific publications through MEDLINE (Airozo et al. 1999) has not only severely affected the ways of writing papers and doing science in general, it has also enabled the development of an avalanche of methods that mine these data. Automatic text-analysis tools can assist human annotators and can thus significantly shorten the time lag of functional annotations. One of the most crucial bottlenecks for automated text analysis is the mapping of gene-protein names (Hatzivassiloglou et al. 2001; Valencia 2002). Although this problem might be overcome in the near future by particular standards adopted by journals, this hurdle currently hinders the availability and usefulness of public methods considerably.

Many tools focus on mining MEDLINE abstracts. Although the principal reason for this restriction is supposedly related to complexity (abstracts available fit onto a disk and can be searched quickly), abstracts are occasionally more easy to mine because many papers contain less precise and less well-supported sections in the text that are difficult for machines to distinguish from more informative sections (Andrade and Bork 2000; Ding et al. 2002; Hersh et al. 1992).

The current version of MEDLINE contains nearly 12 million abstracts and occupies approximately 43 gigabytes of disk space. A prominent example of methods that target entire papers is still restricted to a small number of journals (Friedman et al. 2001; Krauthammer et al. 2002). The task of unraveling information about function from MEDLINE abstracts can be approached from two different angles. In the first approach, computational techniques for understanding text written in natural language are based on lexical, syntactical, and semantic analysis (Cowie and Lehnert 1996; Salton 1989). In addition to indexing terms in documents, natural language–processing methods extract and index higher-level semantic structures composed of terms and relationships between terms, which can be done in different ways (for general discussion, see Baeza-Yates and Ribeiro-Neto [1999]). However, this approach is confronted with the variability, fuzziness, and complexity of human language (Andrade and Bork 2000). The GENIES system (Friedman et al. 2001; Krauthammer et al. 2002) for automatically gathering and processing knowledge about molecular pathways and the IFBP transcription-factor database (Ohta, Yamamoto et al. 1997) are natural language–processing-based systems. The second approach and one that might be more relevant in practice is based on the treatment of text with statistical methods (Andrade and Valencia 1998; Yang 1996). In this approach, the possible relevance of words in a text is deduced from a comparison of the frequency of different words in this text with the frequency of the same words in reference sets of text (Berry et al. 1995). Some of the major methods using the statistical approach are ABXTRACT (Andrade and Valencia 1998, 1997) and the automatic pathway discovery tool of Ng and Wong (1999) (table 2).

There are advantages to each of these approaches (grammar or pattern matching). Generally, the less syntax that is used, the more domain specific the system is. Thus, a robust system can be constructed relatively quickly, but many subtleties can be lost in the interpretation of the sentence. In some applications, however, the domain-dependent pattern-matching approach might be the only way to attain reasonable performance in the near future (Allen 1995).

The ABXTRACT system (Andrade and Valencia 1998, 1997; Blaschke 2001; Blaschke and Valencia 2001) is triggered by collections of abstracts related to a given protein, and it is able to extract functional information directly from MEDLINE abstracts. Relevant keywords are selected by their relative accumulation in comparison with a domain-specific background distribution. To obtain a representative set of words (and their abundance) in protein families, the background distribution of abstracts is chosen to represent the widest range of protein families. For each of the representative set (dictionary) of words, two statistical parameters are computed: (1) their frequency in each family and (2) the deviation of the distribution of their frequencies in the set of families. Provided with a query family and an associated set of MEDLINE abstracts, words that are likely to be functionally important for the family (putative keywords) are found by comparison with the background set. This comparison is done by measuring the frequency of the relevant word in the query family relative to its background frequency of occurrence using a *z*-score (Andrade and Valencia 1998). Words with a high *z*-score are likely to be potential keywords for the family.

The system has been tested on a number of different protein families and showed a good ability to extract functionally important keywords. A modification of this algorithm, called SUISEKI (system for information extraction on interactions) (Blaschke and Valencia 2001; Blaschke et al. 2002), has been applied to the problem of extracting protein-protein interaction from MEDLINE abstracts. In addition to the statistical approach of AbXtract, SUISEKI also takes advantage of the analysis of the syntactic structure of phrases and other developments in computational linguistics. The SUISEKI system was able to extract almost 70 percent of the interactions present in a relatively large text corpus at approximately 80-percent accuracy for the best-defined interactions. The SUISEKI system discovered a total of 4657 protein-protein interactions in cell-cycle proteins in yeast from a corpus of approximately 5300 abstracts (approximately 12 megabytes).

The authors identify a number of sources of error in mining MEDLINE abstracts: First is erroneous detection of protein names. There is no systematic nomenclature for gene and protein names, which has led to a number of possible writing variants and synonyms being associated with the proteins that makes detection and classification difficult. Second are indirect references and anaphoric expression. This problem is key for the analysis of MEDLINE abstracts, where protein names can be given in the title or initial sentences and later treated with forms such as *the protein* or mentioned as a general class of proteins such as *the kinase*. Third are deficiencies in the information-extraction technology. Incorrect parsing of sentences as a result of limitations imposed by the parsers and the use of complicated sentence structures to convey results are some other areas where the information-extraction applications require improvement.

Conclusions

What is to be expected from computational genomics in the near future? As we illustrated in this article, our battery of tools is becoming increasingly sophisticated, and our ability to annotate protein function using computers is generally improving. However, to fully exploit genome information, we still need to progress from methods derived mostly from traditional sequence analysis that examine genome sequences individually to algorithms and databases that exploit the inherent properties of entire genomes. The development of a standardized ontology is an important step in this direction. Text-based tools such as LOCkey, EUCLID, and the Met_A annotator that infer cellular function from detailed annotations of molecular function found in databases can be useful aids in the development of ontologies.

The development of tools for the extraction of useful functional information from the existing literature is still in its infancy. The complexity and fuzziness of natural language make it extremely difficult for computer algorithms to parse and extract useful information from text. The development of a standardized vocabulary for reporting experimental discovery in the scientific literature will go a long way toward simplifying the process of extracting information directly from literature.

Genomics-based drug discovery is heavily dependent on accurate functional annotations. Toward this end, bioinformatics will need to deliver highly integrated, interoperable data “warehouses” that allow the user to reason over disparate data sources and ultimately enable knowledge-based inference and innovation. The road toward satisfactory solutions might be long. However, the first successes have been encouraging. One important lesson from the successes of bioinformatics over the last decade continues to be that integrated tools become successful when their developers are integrated with the wet-lab biologists.

Acknowledgments

Thanks to the following members of our group: Jinfeng Liu and Megan Restuccia, for computer assistance, and Dariusz Przybylski and Kazimierz Wrzeszczynski, for helpful discussions. The authors were supported by grant DBI-0131168 from the National Science Foundation. Last, but not least, thanks to all those

who deposit their experimental data in public databases and to those who maintain these databases.

References

Airozo, D.; Allard, R.; Brylawski, B.; Canese, K.; Kenton, D.; Knecht, L.; Krasnov, S.; Sandomirskiy, V.; Sirotinin, V.; Starchenko, G.; Wilbur, J.; and Zipser, J. 1999. MEDLINE. Bethesda, Md.: National Library of Medicine.

Alberts, B.; Bray, D.; Roberts, K.; and Watson, J. 1994. *Molecular Biology of the Cell.* New York: Garland.

Allen, J. 1995. *Natural Language Understanding.* New York: Addison-Wesley.

Andrade, M. A., and Bork, P. 2000. Automated Extraction of Information in Molecular Biology. *Federation of European Biochemical Societies Letters* 476(1–2): 12–17.

Andrade, M. A., and Valencia, A. 1998. Automatic Extraction of Keywords from Scientific Text: Application to the Knowledge Domain of Protein Families. *Bioinformatics* 14(7): 600–607.

Andrade, M. A., and Valencia, A. 1997. Automatic Annotation for Biological Sequences by Extraction of Keywords from MEDLINE Abstracts. Development of a Prototype System. In *Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology,* 25–32. Menlo Park, Calif.: AAAI Press.

Andrade, M.; Brown, A. N.; Leroy, C.; Hoersch, S.; de Daruvar, A.; Reich, C.; Franchini, A.; Tamames, J.; Valencia, A.; Ouzounis, C.; and Sander, C. 1999. Automated Genome Sequence Analysis and Annotation. *Bioinformatics* 15(5): 391–412.

Andrade, M. A.; Ouzounis, C.; Sander, C.; Tamames, J.; and Valencia, A. 1999. Functional Classes in the Three Domains of Life. *Journal of Molecular Evolution* 49(5): 551–557.

Apte, C.; Damerau, F.; and Weiss, S. 1994. Toward Language-Independent Automated Learning of Text Categorization Models. In Proceedings of the Seventeenth Annual ACM/SIGIR Conference, 24–30. New York: Association of Computing Machinery.

Apweiler, R. 2001. Functional Information in SWISS-PROT: The Basis for Large-Scale Characterization of Protein Sequences. *Briefings in Bioinformatics* 2(1): 9–18.

Apweiler, R.; Gateau, A.; Contrino, S.; Martin, M. J.; Junker, V.; O'Donovan, C.; Lang, F.; Mitaritonna, N.; Kappus, S.; and Bairoch, A. 1997. Protein Sequence Annotation in the Genome Era: The Annotation Concept of SWISS-PROT + TrEMBL. In *Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology,* 33–43. Menlo Park, Calif.: AAAI Press.

Ashburner, M.; Ball, C. A.; Blake, J. A.; Botstein, D.; Butler, H.; Cherry, J. M.; Davis, A. P.; Dolinski, K.; Dwight, S. S.; Eppig, J. T.; Harris, M. A.; Hill, D. P.; Issel-Tarver, L.; Kasarskis, A.; Lewis, S.; Matese, J. C.; Richardson, J. E.; Ringwald, M.; Rubin, G. M.; and Sherlock, G. 2000. Gene Ontology: Tool for the Unification of Biology. The Gene Ontology Consortium. *Nature Genetics* 25(1): 25–29.

Attwood, T. K. 2000. Genomics. The Babel of Bioinformatics. *Science* 290(5491): 471–473.

Baeza-Yates, R.; and Ribeiro-Neto, B., eds. 1999. *Modern Information Retrieval.* New York: ACM Press.

Bairoch, A., and Apweiler, R. 2000. The SWISS-PROT Protein Sequence Database and Its Supplement TrEMBL in 2000." *Nucleic Acids Research* 28(1): 45–48.

Baker, P. G., and Brass, A. 1998. Recent Developments in Biological Sequence Databases. *Current Opinion in Biotechnology* 9(1): 54–58.

Bazzan, A. L.; Engel, P. M.; Schroeder, L. F.; and Da Silva, S. C. 2002. Automated Annotation of Keywords for Proteins Related to Mycoplasmataceae Using Machine Learning Techniques. *Bioinformatics* 18(Suppl. 2): S35–S43.

Berry, M. W.; Dumais, S. T.; and O'Brien, G. W. 1995. Using Linear Algebra for Intelligent Information Retrieval. *Society for Industrial and Applied Mathematics Review* 37(4): 573–595.

Blaschke, C. 2001. Applications of Information-Extraction Techniques to Molecular Biology. Ph.D. diss., CNB, Madrid, University of Autonoma Madrid.

Blaschke, C., and Valencia, A. 2002. The Frame-Based Module of the SUISEKI Information-Extraction System. *IEEE Intelligent Systems* 17(1): 14–20.

Blaschke, C., and Valencia, A. 2001. The Potential Use of SUISEKI as a Protein Interaction Discovery Tool. Genome Informatics Series Workshop. *Genome Informatics* 12:123–134.

Blaschke, C.; Hirschman, L.; and Valencia, A. 2002. Information Extraction in Molecular Biology. *Briefings in Bioinformatics* 3(2): 154–165.

Boeckmann, B.; Bairoch, A.; Apweiler, R.; Blatter, M. C.; Estreicher, A.; Gasteriger, E.; Martin, M. J.; Michoud, K.; O'Donovan, C.; Phan, I.; Pibout, S.; and Schneider, M. 2003. The SWISS-PROT Protein Knowledge Base and Its Supplement TrEMBL in 2003. *Nucleic Acids Research* 31(1): 365–370.

Bork, P.; Dandekar, T.; Diaz-Lazcoz, Y.; Eisenhaber, F.; Huynen, M.; and Yuan, Y. 1998. Predicting Function: From Genes to Genomes and Back. *Journal of Molecular Biology* 283(4): 707–725.

Bork, P,. and Koonin, E. V. 1998. Predicting Functions from Protein Sequences—Where Are the Bottlenecks? *Nature Genetics* 18(4): 313–318.

Bork, P.; Ouzounis, C.; Sander, C.; Scharf, M.; Schneider, R.; and Sunnhammer, E. 1992. What's in a Genome?" *Nature* 358(6384): 287.

Brutlag, D. L. 1998. Genomics and Computational Molecular Biology. *Current Opinion in Microbiology* 1(3): 340–345.

Carter, P.; Liu, J.; and Rost, B. 2003. PEP: Predictions for Entire Proteomes. *Nucleic Acids Research* 31(1): 410–413.

Cowie, J., and Lehnert, W. 1996. Information Extraction. *Communications of the ACM* 39(1): 80–91.

Dasarathy, B. V. 1991. Nearest-Neighbor (NN) Norms: NN Pattern Classification Techniques. Washington, D.C.: IEEE Computer Society.

Devos, D., and Valencia, A. 2001. Intrinsic Errors in Genome Annotation. *Trends in Genetics* 17(8): 429–431.

Devos, D., and Valencia, A. 2000. Practical Limits of Function Prediction. *Proteins* 41(1): 98–107.

Ding, J.; Berleant, D.; Nettleton, D.; and Wurtele, E. 2002. Mining MEDLINE: Abstracts, Sentences, or Phrases? Paper presented at the Pacific Symposium on Biocomputing, 6–10 January, Kapalua, Hawaii.

Doerks, T.; Bairoch, A.; and Bork, P. 1998. Protein Annotation: Detective Work for Function Prediction. *Trends in Genetics* 14(6): 248–250.

Eisenberg, D.; Marcotte, E. M.; Xenarios, I.; and Yeates, T. O. 2000. Protein Function in the Post-Genomic Era. *Nature* 405(6788): 823–826.

Eisenhaber, F., and Bork, P. 1999. Evaluation of Human-Readable Annotation in Biomolecular Sequence Databases with Biological Rule Libraries. *Bioinformatics* 15(7–8): 528–535.

Fleischmann, R. D.; Adams, M. D.; White, O.; Clayton, R. A.; Smith, H. O.; and Venter, J. C. 1995. Whole-Genome Random Sequencing and Assembly of Haemophilus Influenzae Rd. *Science* 269(5223): 496–512.

Fleischmann, W.; Moller, S.; Gateau, A.; and Apweiler, R. 1999. A Novel Method for Automatic Functional Annotation of Proteins. *Bioinformatics* 15(3): 228–233.

Fraser, C. M.; Gocayne, J. D.; White, O.; Adams, M. D.; Philips, C. A.; and Venter, J. C. 1995. The Minimal Gene Complement of Mycoplasma Genitalium. *Science* 270(5235): 397–403.

Friedman, C.; Kra, P.; Yu, H.; Krauthammer, M.; Rzhetsky, A. 2001. GENIES: A Natural Language–Processing System for the Extraction of Molecular Pathways from Journal Articles. *Bioinformatics* 17(Suppl. 1): S74–S82.

Gaasterland, T., and Sensen, C. W. 1996. MAGPIE: Automated Genome Interpretation. *Trends in Genetics*12(2): 76–78.

Galperin, M. Y., and Koonin, E. V. 2000. Who's Your Neighbor? New Computational Approaches for Functional Genomics. *Nature Biotechnology* 18(6): 609–613.

Harrison, P. M.; Bamborough, P.; Daggett, V.; Prusiner, S.; and Cohen, F. E. 1997. The Prion Folding Problem. *Current Opinion in Structural Biology* 7(1): 53–59.

Hatzivassiloglou, V.; Duboue, P. A.; and Rzhetsky, A. 2001. Disambiguating Proteins, Genes, and RNA in Text: A Machine Learning Approach. *Bioinformatics* 17(Suppl. 1): S97–S106.

Hersh, W. R.; Evans, D. A.; Monarch, I. A.; Lefferts, R. G.; Handerson, S. K.; and Gorman, P. N. 1992. *Indexing Effectiveness of Linguistic and Nonlinguistic Approaches to Automatic Indexing*. Amsterdam, The Netherlands: Elsevier Science.

Iliopoulos, I.; Enright, D. A.; and Ouzounis, C. A. 2001. TEXTQUEST: Document Clustering of MEDLINE Abstracts for Concept Discovery in Molecular Biology. Paper presented at the Pacific Symposium on Biocomputing, 3–7 January, Mauna Lani, Hawaii.

Junker, V.; Contrino, S.; Reischmann, W.; and Apweiler, R. 2000. The Role SWISS-PROT and TrEMBL Play in the Genome Research Environment. *Journal of Biotechnology* 78(3): 221–234.

Karp, P. D.; Riley, M.; Paley, S. M.; Pellegrini-Toole, A.; and Krummenacker, M. 1999. Eco CYC: Encyclopedia of *Escherichia coli* Genes and Metabolism. *Nucleic Acids Research* 27(1): 55–58.

Koonin, E. V. 2000. Bridging the Gap between Sequence and Function. *Trends in Genetics* 16(1): 16.

Krauthammer, M.; Kra, P.; Iossifov, I.; Gomez, S. M.; Hripcsak, G.; Hatzivassiloglou, V.; Friedman, C.; and Rzhetsky, A.. 2002. Of Truth and Pathways: Chasing Bits of Information through Myriads of Articles. *Bioinformatics* 18(Suppl. 1): S249–S257.

Krawiec, S., and Riley, M. 1990. Organization of the Bacterial Chromosome. *Microbiology Review* 54(4): 502–539.

Kretschmann, E.; Fleischmann, W.; et al. 2001. Automatic Rule Generation for Protein Annotation with the C4.5 Data-Mining Algorithm Applied on SWISS-PROT. *Bioinformatics* 17(10): 920–926.

Lewis, D. D., and Ringuette, M. 1994. Comparison of Two Learning Algorithms for Text Categorization. Paper presented at the Third Annual Symposium on Document Analysis and Information Retrieval (SDAIR'94), 11–13 April, Las Vegas, Nevada.

Lewis, S.; Ashburner, M.; and Reese, M. G. 2000. Annotating Eukaryote Genomes. *Current Opinion in Structural Biology* 10(3): 349–354.

Liu, J., and Rost, B. 2001. Comparing Function and Structure between Entire Proteomes. *Protein Science* 10(10): 1970–1979.

Lodish, H.; Berk, A.; Baltimore, D.; and Darnell, J. 2000. *Molecular Cell Biology.* New York: Freeman.

Luscombe, N. M.; Greenbaum, D.; and Gerstein, M. 2001. What Is Bioinformatics? A Proposed Definition and Overview of the Field. *Methods of Information in Medicine* 40(4): 346–358.

Mewes, H. W.; Frishman, D.; Gruber, C.; Stocker, S.; and Weil, B. 2000. MIPS: A Database for Genomes and Protein Sequences. *Nucleic Acids Research* 28(1): 37–40.

Nair, R., and Rost, B. 2002a. Inferring Subcellular Localization through Automated Lexical Analysis. *Bioinformatics* 18(Suppl. 1): S78–S86.

Nair, R., and Rost, B. 2002b. Sequence Conserved for Subcellular Localization. *Protein Science* 11(12): 2836–2847.

Ng, S. K., and Wong, M. 1999. Toward Routine Automatic Pathway Discovery from Online Scientific Text Abstracts. Genome Informatics Series Workshop. *Genome Informatics* 10(1): 104–112.

Ohta, Y.; Yamamoto, Y.; Okazachi, T.; Uchiyama, I.; and Takagi, T. 1997. Automatic Construction of Knowledge Base from Biological Papers. Paper presented at the International Conference of Intelligent Systems in Molecular Biology, 21–25 June, Halkidiki, Greece.

Ouzounis, C.; Casari, G.; Sander, C.; Tamames, J.; and Valencia, A. 1996. Computational Comparisons of Model Genomes. *Trends in Biotechnology* 14(B): 280–285.

Ouzounis, C.; Perez-Irratxeta, C.; Sander, C.; and Valencia, A. 1998. Are Binding Residues Conserved? Paper presented at the Pacific Symposium on Biocomputing, 4–9 January, Kapalua, Maui, Hawaii.

Overbeek, R.; Larsen, N.; Smith, W.; Maltsev, N.; and Selkov, E. 1997. Representation of Function: The Next Step. *Gene* 191(1): GC1–GC9.

Pruess, M.; Fleischmann, W.; Kanapin, A.; Servant, F.; and Apweiler, R. 2003. The Proteome Analysis Database: A Tool for the In Silico Analysis of Whole Proteomes. *Nucleic Acids Research* 31(1): 414–417.

Riley, M. 1993. Functions of the Gene Products of *Escherichia coli*. *Microbiology Review* 57(4): 862–952.

Riley, M., and Labedan, B. 1997. Protein Evolution Viewed through *Escherichia coli* Protein Sequences: Introducing the Notion of a Structural Segment of Homology, the Module. *Journal of Molecular Biology* 268(5): 857–868.

Rost, B. 2002. Enzyme Function Less Conserved Than Anticipated. *Journal of Molecular Biology* 318(2): 595–608.

Rost, B. 1999. Twilight Zone of Protein Sequence Alignments. *Protein Engineering* 12(2): 85–94.

Rost, B., and Sander, C. 1996. Bridging the Protein Sequence-Structure Gap by Structure Predictions. *Annual Review of Biophysics and Biomolecular Structure* 25:113–136.

Salton, G. 1989. *Automatic Text Processing*. Reading, Mass.: Addison-Wesley.

Schutze, H.; Hull, D. A.; and Pederson, J. O. 1995. A Comparison of Classifiers and Document Representation for the Routing Problem. In Proceedings of the Ann Int ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '95), 229–237. New York: Association of Computing Machinery.

Shah, I., and Hunter, L. 1997. Predicting Enzyme Function from Sequence: A Systematic Appraisal. In *Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology,* 276–283. Menlo Park, Calif.: AAAI Press.

Shatkay, H.; Edwards, S.; William, W. J.; and Boguski, M. 2000. Genes, Themes, and Microarrays: Using Information Retrieval for Large-Scale Gene Analysis. Paper presented at the International Conference on Intelligent Systems in Molecular Biology, 19–23 August, La Jolla, California.

Stapley, B. J., and Benoit, G. 2000. Biobibliometrics: Information Retrieval and Visualization from Cooccurrences of Gene Names in Medline Abstracts. Paper presented at the Pacific Symposium on Biocomputing, 4–8 January, Honolulu, Hawaii.

Stapley, B. J.; Kelley, L. A.; and Sternberg, M. J. 2002. Predicting the Subcellular Location of Proteins from Text Using Support Vector Machines. Paper presented at the Pacific Symposium on Biocomputing, 3–7 January, Lihue, Hawaii.

Stephens, M.; Palakal, M.; Mukhopadhyay, S.; Raje, R.; and Mostata, J. 2001. Detecting Gene Relations from MEDLINE Abstracts. Paper presented at the Pacific Symposium on Biocomputing, 3–7 January, Honolulu, Hawaii.

Stevens, R.; Goble, C. A.; and Bechhofer, S. 2000. Ontology-Based Knowledge Representation for Bioinformatics. *Briefings in Bioinformatics* 1(4): 398–414.

Tamames, J.; Ouzounis, C.; Casari, G.; Sander, C.; and Valencia, A. 1998. EUCLID: Automatic Classification of Proteins in Functional Classes by Their Database Annotations. *Bioinformatics* 14(6): 542–543.

Thomas, J.; Milward, D.; Ouzounis, C.; Pulman, S.; and Carroll, M. 2000. Automatic Extraction of Protein Interactions from Scientific Abstracts. Paper presented at the Pacific Symposium on Biocomputing, 4–8 January, Honolulu, Hawaii.

Todd, A. E.; Orengo, C. A.; and Thornton, J. M. 2001. Evolution of Function in Protein Superfamilies, from a Structural Perspective. *Journal of Molecular Biology* 307(4): 1113–1143.

Tsoka, S., and Ouzounis, C. A. 2000. Recent Developments and Future Directions in Computational Genomics. *Federation of European Biochemical Societies Letters* 480(1): 42–48.

Valencia, A. 2002. Search and Retrieve: Large-Scale Data Generation Is Becoming Increasingly Important in Biological Research. But How Good Are the Tools to Make Sense of the Data? *European Molecular Biology Organization (EMBO) Reports* 3(5): 396–400.

Valencia, A., and Pazos, F. 2002. Computational Methods for the Prediction of Protein Interactions. *Current Opinion in Structural Biology* 12(3): 368–373.

Webb, E. C. 1992. *Enzyme Nomenclature 1992. Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology.* San Diego, Calif.: Academic.

Wilson, C. A.; Kreychman, J.; and Gerstein, M. 2000. Assessing Annotation Transfer for Genomics: Quantifying the Relations between Protein Sequence, Structure, and Function through Traditional and Probabilistic Scores. *Journal of Molecular Biology* 297(1): 233–249.

Wrzeszczynski, K. O., and Rost, B. 2003. Cataloguing Proteins in Cell Cycle Control. In *Cell Cycle Checkpoint Control Protocols,* ed. H. Lieberman, 219–233. Totowa, N.J.: Humana.

Yakushiji, A.; Tateisi, Y.; Miyao, Y.; and Tsujii, J. 2001. Event Extraction from Biomedical Papers Using a Full Parser. Paper presented at the Pacific Symposium on Biocomputing, 3–7 January, Honolulu, Hawaii.

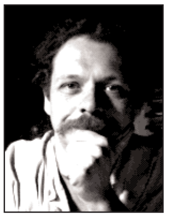
Yang, Y. 1996. An Evaluation of Statistical Approaches to MEDLINE Indexing. Paper presented at the Conference of the American Medical Informatics Association, 26–29 October, Washington, D.C.

Yang, Y., and Chute, C. G. 1992. An Application of Least Squares Fit Mapping to Clinical Classification. Paper presented at the Annual Symposium on Computer Applications in Medical Care, 2–6 November, Washington, D.C.

Yang, Y., and Liu, X. 1999. A Reexamination of Text Categorization Methods. In Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99), 42–49. New York: Association of Computing Machinery.

Yang, Y., and Pederson, J. P. 1997. A Comparative Study on Feature Selection in Text Categorization. Paper presented at the Fourteenth International Conference on Machine Learning, 8–12 July, Nashville, Tennessee.

Rajesh Nair is a graduate student in the Physics Department at Columbia University. He has been responsible for developing a number of widely used bioinformatics tools for the prediction of protein subcellular localization such as PredictNLS, LOCkey, LOChom, and LOC3d. His recent research interests have focused on the application of machine learning techniques such as neural networks and support vector machines to predicting various aspects of protein function. His e-mail address is nair@cubic.bioc.columbia.edu.



Burkhard Rost is an associate professor of biochemistry and molecular biophysics at Columbia University. He was responsible for creating the PREDICT PROTEIN server for protein secondary structure prediction. PREDICT PROTEIN is the most widely used public server for protein structure prediction and has handled over a million requests. His research focuses on the prediction of protein structure and function by combining means from simple statistics to AI and evolutionary information. His e-mail address is rost@columbia.edu.