An integrative algorithmic approach towards knowledge discovery by bioinformatics

Blaise T.F. Alako

Promotor:	Prof. dr. J.A.M Leunissen Hoogleraar Bioinformatica Laboratorium voor Bioinformatica Wageningen Universiteit
Promotiecommissie:	Prof. dr. G. Vriend (Radboud Universiteit, Nijmegen) Dr. ir. G.W. Jenster (Erasmus Universiteit Rotterdam) Prof. dr. M. Groenen (Wageningen Universiteit) Prof. dr. A.H.J. Bisseling (Wageningen Universiteit)

Dit onderzoek is uitgevoerd binnen de onderzoekschool "Experimental Plant Sciences"

An integrative algorithmic approach towards knowledge discovery by bioinformatics

Blaise T.F. Alako

Proefschrift
ter verkrijging van de graad van doctor
op gezag van de rector magnificus
van Wageningen Universiteit,
prof. dr. M.J. Kropff,
in het openbaar te verdedigen
op vrijdag 1 februari 2008
des namiddags te 16:00 uur in de Aula

Blaise T.F. Alako, 2008
An integrative algorithmic approach towards knowledge discovery by bioinformatics

PhD thesis Wageningen University, The Netherlands
With summaries in English and Dutch

ISBN 978-90-8504-819-0

This book is dedicated with greatest love and affection to:
My dear parents Alako Etienne and Assongou Victoire
My sister Likeufack hortense
My godfather Ndeh Christopher
My wife Alako Liliane
Our daughter Alako Kaycee
Our expected child Alako ...

Abstracts

Title:

An integrative algorithmic approach towards knowledge discovery by bioinformatics

Short Summary:

In this thesis we describe different approaches aiding in the utilization of the exponentially growing amount of information available in the life sciences. Briefly, we address two issues in molecular biology, on sequence analysis, and on text mining. The former issue addresses the problem how to determine remote sequence homology especially when the sequence similarity is very low. For this a visualisation tool is introduced that combines sequence alignment, domain prediction and phylogeny. The second topic on text mining centres on the question how to unambiguously formulate queries for efficient information retrieval. It tackles the problem of gene nomenclature – one in two gene symbols being ambiguous - by introducing a new text-clustering- and taxonomy-based disambiguation methodology.

Titel: (Title in Dutch)

Kennisextractie ("knowledge discovery") in de moleculaire biologie met behulp van bioinformatica technieken

Samenvatting: (Short summary in Dutch)

In dit proefschrift worden verschillende benaderingen beschreven om de exponentieel groeiende data stroom in de levenswetenschappen te exploiteren. Twee velden uit de moleculaire biologie worden aangesproken, de sequentie analyse en tekst analyse.
Het eerste onderwerp richt zich op het probleem van de verwantschap van eiwitten, vooral wanneer de onderlinge overeenkomsten erg gering zijn. Hiervoor is een visualisatie programma ontwikkeld dat sequentie alignment, eiwit domein voorspelling en fylogenie combineert.
Het tweede hoofdonderwerp van dit proefschrift, text mining, spreekt het probleem aan hoe een zoekvraag in de literatuur zo efficiënt en eenduidig mogelijk geformuleerd kan worden. Hierbij wordt in het bijzonder het probleem van de gen-symbool ambiguïteit geadresseerd: 50% van de gen-symbolen is niet eenduidig, en kan dus meer dan een gen of eiwit familie aanwijzen. De nieuw ontwikkelde methode kan door gebruik te maken van tekst clustering en taxonomische informatie dit probleem in vrijwel alle gevallen eenduidig oplossen.

Table of Contents

Chapter 1

General introduction

1.1 Definition of bioinformatics

Bioinformatics, often referred to as computational biology, involves the use of various fields including mathematics, statistics, computer science, artificial intelligence, chemistry and biochemistry to solve biological problems, most commonly at the molecular level. Major bioinformatics research interests comprise diverse domains such as sequence alignment, gene finding, genome assembly, protein structure prediction, protein-protein interaction, phylogeny, text mining and many more. Whereas computational biology is focussed on forming a hypothesis based on a given data set, bioinformatics is more concerned with gaining information from such data. In other words, bioinformatics refers more to the creation and advancement of algorithms, computational and statistical techniques and theory to solve formal and practical problems deduced from the management and analysis of biological data. Meanwhile, computational biology refers to a hypothesis-driven investigation of a specific biological problem using computers as tools. The latter is carried out with experimental or simulated data, with the primary goal of discovery and the advancement of biological knowledge.

1.2 Why using bioinformatics?

Nowadays genome-wide techniques such as micro array analysis, Serial Analysis of Gene Expression (SAGE), Massively Parallel Signature Sequencing (MPSS), linkage analysis, yeast two-hybrid, mass spectrometry and association studies are used extensively in the search for genes that are causative in diseases or responsible for a phenotype in general, and these techniques often identify many hundreds of candidate genes. Such high-throughput experimental technologies have given rise to the "omics" fields in current life sciences that are characterized by the transition from local to global scale studies, as well as generating complete genomic sequences of unprecedented number and size.
In view of such massive generation of information new approaches are required to aid in the organization and exploitation of these data sets in order to live up to the expectations that the "omics" fields may rise.

1.3 Computational technique and machine learning in bioinformatics

The exponential growth of biological data raises two main problems: on one hand the efficient storage and management, on the other hand the extraction of useful information from these data. The latter requires the development of tools and methods to transform these heterogeneous data into biologically meaningful facts and testable models.
The ultimate goal is to understand and predict normal function of organisms and to proceed from there to the understanding of abnormalities such as diseases. Given the wealth of data, the interpretation can not be done manually. It requires advanced computational tools, mimicking some aspects of the manual interpretation process and thus computer science becomes an indispensable asset in assisting the automation of data analysis in biology.
Machine learning (ML) is a field of research where computational methods learn to answer complicated problems based on sets of provided data. Machine learning algorithms are data-driven and are ideally suited for areas with an extensive generation of data but little theoretical background, such as is often the case in the field of molecular biology. The methods often do not need to be separately modified for each problem; rather they are general-purpose. Classifying samples, pattern recognition, clustering, modelling, and visualization are typical applications of machine learning. Machine

learning can be largely defined as either supervised or deductive methods that attempt to obtain a "correct response" given by a teacher, unsupervised or inductive methods that attempt to achieve the statistical goal "response unknown", and reinforcement learning where one is rewarded based on how well one did, but is not told what the correct response was. Machine learning uses computer programs to optimize a performance criterion by managing example data or past experiences. In other words, machine learning is concerned with the design and development of algorithms and techniques that allow computers to "learn". The ultimate goal here is to extract useful information from a body of data by building good probabilistic models and automating the process as much as possible. As stated above, in general there are two types of learning: inductive and deductive. Inductive machine learning methods extract rules and patterns out of massive data sets whereas deductive machine learning methods apply the kind of reasoning where drawing a conclusion is necessitated by previously known premises. There are several sub-disciplines of the life sciences where machine learning is applied for knowledge extraction from data. Examples of ML in bioinformatics include Support Vector Machines (SVM), Nearest Neighbour Algorithms (NNA) and Hidden Markov Models (HMM) to name a few. Support vector machines map input vectors to a higher dimensional space where a maximal separating hyperplane is constructed. Two parallel hyperplanes are constructed on each side of the hyperplane that separates the data. The separating hyperplane is the hyperplane that maximizes the distance between the two parallel hyperplanes. An assumption is made that the larger the margin or distance between these parallel hyperplanes the better the generalisation error of the classifier will be. In NNA an object is classified by a majority vote of its neighbours, with the object being assigned to the class most common amongst its k nearest neighbours. Hence NNA is a method for classifying objects based on closest training example in feature space. Features are the individual measurable heuristic properties of the phenomena being observed. A HMM on the other hand is a statistical model in which the system being modelled is assumed to be a Markov process with unknown parameters, and the challenge is to determine the hidden parameters from the observable parameters. The extracted model parameters can be further used for pattern recognition application.

1.4 Bioinformatics in various biological fields

Current ***genomics*** deals with an exponential sequence data production that encompasses for example linkage analysis, sequence assembly, gene annotation, and single nucleotide polymorphism (SNP).
Bioinformatics, using genome sequences, can aid in their analysis, such as gene finding or RNA gene finding, alternative splicing, coding-region identification and splice site prediction as well as the prediction of gene function and RNA secondary structure. Computational techniques can further assist amongst others in motif detection of transcription binding sites, promoter binding sites and operons.
Examples of structure and function prediction by machine learning include the Hidden Markov Model (HMM), multilayer perceptrons and decision trees. A perceptron is a binary classifier that maps its input (binary vector) to an output single binary value. A decision tree is a predictive model that uses graph decisions to map observation about an item to conclusion about its target value. Decision trees are also known as classification trees (discrete outcome) and regression trees (continuous outcome).

In the ***proteomics*** domain, the main application of machine learning can be found in protein function and structure prediction, protein location prediction, protein-protein interaction and protein annotation as well as sequence alignment (Altschul, Madden et al.

1997; Notredame, Higgins et al. 2000; Edgar 2004; Ma, G et al. 2007) (simulated annealing, genetic algorithm).

Bioinformatics also finds its application in the management of complex ***microarray*** experimental data sets, which has lead to the application of techniques such as Self Organising Map (SOM) and Bayesian network. Microarray assay's complexity owes to the fact that data have to be pre-processed, i.e. modified, prior to their analysis by machine learning algorithms in search for expression pattern identification and genetic network identification. Traditional clustering techniques based on hierarchical, self-organising maps (Tamayo and et al. 1988) have been used to derive putative functional clusters of genes from expression profile data (Sherlock 2001) . Applying the guilt-by-association principle, expression profile clustering also can be used for inferring the biological functions of new genes. If an uncharacterised gene is clustered with a group of genes known to participate in a specific biological process, then it is hypothesized that the uncharacterised gene also participates in this process.
However, genes sharing similar expression profiles do not always share a common function. Spellman and colleagues (Spellman and et al. 1988) reported that clustering by expression profile grouped genes into a single cluster even though they are involved in distinct cellular functions. The reverse can also hold true, i.e. not all genes comprised in the same function group necessarily exhibit a common expression. For example, the members of a signalling pathway often play antagonistic roles, resulting in anticorrelated expression levels in microarray experiments. Therefore the gene expression clustering approach should not be used as the sole analysis tool, but rather should be coupled with other data mining techniques. The latter can provide necessary biological knowledge in intelligent expression profile analysis.

Another domain of biology where machine learning proves valuable is ***systems biology***. The objective of systems biology is to model intracellular life processes, cell, organ, and organism up to whole ecosystems. Bioinformatics uses computational techniques to model these biological networks such as metabolic pathways, genetic and signal transduction networks.

The field of ***evolution*** attempts to elucidate and understand the changes of inherited traits in a population from generation to generation. An important technique in this field is the generation of phylogenetic trees as the schematic representation of organism's evolution. These trees are currently based on the comparison of different genomes (molecular evolution), a comparison that is made based on multiple sequence alignment where computational techniques are extremely valuable for their optimization (Baldi P and S. 2001).

A consequence of the use of computational techniques in the aforementioned biological fields is a dramatic increase in available publications. This in turn represents a new source of valuable information where ***text mining*** techniques are required to keep at pace with the information load. Text mining sees its benefits in for instance functional annotation, cellular location prediction and protein-protein interaction analysis and extraction (Krallinger, Erhardt et al. 2005). It aims to automatically distil information, extract facts, discover implicit links and generate hypothesis relevant to the user's needs (Spasic, Ananiadou et al. 2005). Researchers in life sciences usually design experiments based on prior knowledge embedded in the literature to generate hypotheses that can be experimentally validated or rejected in the laboratory. The available information in literature is however in one of the most challenging formats for large-scale exploitation,

i.e. natural language text. In order to purposefully exploit this wealth of information in scientific publications several tools have been developed to extract and mine data in the literature (Smalheiser and Swanson 1998; Tanabe, Scherf et al. 1999; Liu, Jenssen et al. 2004; Tu, Tang et al. 2004; Zhou, Wen et al. 2004; Alako, Veldhoven et al. 2005).

These developments clearly show the importance of text mining in building on existing knowledge and linking various biological fields to provide a sound basis of directed and concerted actions in current life sciences. Therefore the next section will introduce text mining in more detail as the main scope of this thesis.

1.5 Biological terminology and ontology

In order to share the vast amount of biological and biomedical knowledge effectively, textual evidence needs to be linked to ontologies as the main repositories of formally represented knowledge. These ontological repositories are created by organizing sets of terms in order to define their relationships to each other, thus formalizing knowledge and making it more accessible. In view of this the next logical step is to transform biology into a machine-readable depiction of life as we know it.

Therefore ontologies are crucial for text mining because it provides semantic interpretation of texts and also constrains the possible interpretation of biological entities (terms). Consequently, text mining can be classified as ontology-based or ontology-driven. This means that text mining is used to enrich the ontology and ontology is used to help text mining. Thus: does text mining needs ontology or does ontology needs text mining? The fact is that there is a perpetual circle in which text mining and ontology benefit from each other.

Besides the well-structured ontologies other, less well-organized term collection systems exist including taxonomies, controlled vocabulary, thesauri and dictionaries. Differences among the latter are subtle and are generally collectively referred to as terminologies. In the following we will give some examples on current often-used terminologies and ontologies.
Curators at the National Library of Medicine (NLM) review each article entering MEDLINE, then select terms from the Medline Subject Heading (MeSH) hierarchy that best capture the aim of the article, thereby summarizing it using a controlled vocabulary. The MeSH terminology displays certain coverage of how proteins function in cellular systems, but is by far not exhaustive.
The Gene Ontology (GO) (2006) more accurately captures what happens at the cellular and molecular level. Just as MeSH terms are assigned to individual scientific articles to describe their content, GO terms are assigned to proteins to illustrate what they do, i.e. their molecular function (MF); where they do it, i.e. their cellular component (CC), and to what end, i.e. their biological processes (BP).
MeSH and GO have been merged with other terminologies in the Unified Medical Language System (UMLS) (Bodenreider 2004) . Besides MeSH and GO, other current terminologies and ontologies are summarised in table 1.

Ontology/Terminology	Definition	Annotation
SNOMED	The Systematized Nomenclature of Medicine (Bodenreider 2004)	clinical phenomena
OMIM	The Online Mendelian Inheritance in Man (Hamosh, Scott et al. 2005)	genetic disorders
NCBI taxonomy	The National Center for Biotechnology Information Taxonomy (Wheeler, Barrett et al. 2006)	species
MIPS FunCat	The Munich Information Centre for Protein Sequence Functional Catalogue (Ruepp, Doudieu et al. 2006)	proteins
UWDA	The University of Washington Digital Anatomist (Bodenreider 2004)	parameters related to anatomy
COGs	The Cluster of Orthologous Groups of Protein (Tatusov, Fedorova et al. 2003)	orthologous protein groups
IUBMB	The International Union of Biochemistry and Molecular Biology Enzyme Commission	enzymes
NCI	The National Cancer Institute thesaurus (Bodenreider 2004)	oncological phenomena

Table 1: Ontologies and terminologies used by Bioinformatics

These ontologies and terminologies can help to reliably identify protein names, diseases, biological entities, phenotypes and genotypes in literature with a wide range of applications, such as microarray analysis, biomarker discovery and database curation. Computationally accessible annotation systems such as GO enable one to ask, for instance, if physical interaction significantly correlates with molecular function, subcellular localization, or biological processes, thereby laying the groundwork for better algorithms to add predicted annotation to uncharacterized proteins. Text mining tools have started by relying on ontology and terminology, and the latter promises improvement in term of precision and recall (Doms and Schroeder 2005).

1.6 Gene nomenclature standardization chaos

Despite this wealth of information-rich systems, there are a number of fundamental difficulties when performing text mining. A researcher who is starting to work on a gene or protein (here we use genes and proteins interchangeably) so far unknown to him should consider several databases for obtaining the most relevant information. Particularly he or she should use all available gene names when doing literature search. Terms such as gene or proteins names, drug, chemical compound, and other biological entities are biological objects of primary importance for understanding biochemical processes and therefore are the backbone of scientific communication as they are used to identify domain concepts. Successful term identification is therefore the key to getting access to the stored literature information. But what are those biological terms, how are they created, and to what standard do they comply?

The main barriers to successful term identification are extensive lexical variation, which prevents some terms of being recognized in free text, term synonymy and term homonymy, the latter creating uncertainties with respect to the term's exact identity. Furthermore the biological field is mined with constantly changing and expanding terminology and even more importantly the often-encountered lack of stringent naming convention.
Each gene or protein typically has several names and abbreviations which consequently can lead to so-called term ambiguity. For instance 'Cdc28' is also known as 'cyclin-dependent kinase 1' or 'Cdk1' and, to complicate matters even more, some terms

associated with 'Cdc28' are common English words (e.g. 'hairy root'), biological terms (for example 'SDS') or even names of other genes, such as 'Cdc2' that refers to two completely unrelated genes in budding and fission yeast. Another example is the symbol PSA with a multitude of very different associated long terms: Prostate Specific Antigen, Puromycin-Sensitive Aminopeptidase, Psoriatric Arthritis, Pig Serum Albumin, or one of the more than 100 other meanings of PSA that can be found in the literature (Weeber, Schijvenaars et al. 2003).

It was shown that the highest degree of ambiguity with common English words exists for fly (*Drosophila*). This is due to the frequent phenotypic descriptions that are used as gene names and abbreviations thereof. For example in FlyBase (a database for *Drosophila* molecular genetics), "*WE*" is the abbreviation for a gene named "*Washes eye*". This illustrates that the long form as well as the short form of the gene name are perfect English terms. The gene nomenclature guideline for FlyBase is relatively unrestricted (Tuason, Chen et al. 2004), stating that "gene names must be concise, should allude to the genes function, mutant phenotype or other relevant characteristic, and gene name must be unique and not have been previously used for a *Drosophila* gene, moreover gene names should be inoffensive". This is a rather loose guideline, as no format is proposed for the symbols, and no restrictions about ambiguity with English words or other terms are made. The guideline additionally favours the use of descriptive name, which might be useful for an immediate functional classification of genes by a researcher when reading scientific articles, but clearly results in significant disadvantages for literature search and automatic text processing. Moreover, these non-stringent nomenclature favours ambiguity with English words. Examples of ambiguous gene names in fruit fly are: "*cheap date*", mutants that are especially sensitive to alcohol; interestingly another name for this gene is "*amnesiac*" as the mutant also shows poor memory; "*fruity*", mutants that are not interested in females; "*out cold*", with falling temperature the mutants loose their coordination and eventually paralyze; "*sarah*", mutant flies that are practically sterile; "*van gogh*", swirling wing hair patterns in the mutants resemble the brush strokes in van Gogh's paintings; "*clown*", the clown flies' eyes have a characteristic white and red appearance; "*technical knockout*", the gene is involved in protein transport; "*swiss cheese*", mutant flies' brains have swiss-cheese-like holes.

In contrast to the above, the nomenclature guidelines for the mouse genome database (MGD) and the rat genome database (RGD) explicitly state that "genes that are recognizable orthologs of already-named human genes should be given the same name and symbol as the human gene". The Human genome organization (HUGO) also states "that homologous genes in different vertebrate species should where possible have the same gene nomenclature" and that "the agreement between human and mouse gene nomenclature for many homologous genes should be continued and extended to other species where possible". Generally, the nomenclature of, human, mouse, and rat genes are coordinated with each other by the corresponding committees. This enforces a mapping between orthologs by cross-references, co-assignment of nomenclatures to orthologs genes and thus an increasing unification of the individual nomenclatures.
However, even these more stringent databases are not as free of ambiguity as one would wish. Gene names that show ambiguity with the general English language in humans include: "*hip*" and "*hop*", whose gene products help other proteins to fold correctly, and "*jack-1*", janus kinase, which contains two phosphate-transferring domains. Thus, it got its name from the Roman two-faced gatekeeper of heaven Janus. The abbreviation "JAK" is also said to stand for "just another kinase" as there are so overwhelmingly many kinases in the body that makes it difficult to remember all of them. A "*tigger*" is a

transposable element in the human genome, i.e. it can jump to another location in the genome; "*hedgehog*", a gene first found in fruit fly and named because of the resemblance of its mutant larvae with a hedgehog. One of the human hedgehog genes was named "*sonic hedgehog*" according to the Sega computer game character; "*pokemon*", (POK Erythroid Myeloid Ontogenic Factor) is an oncogene, that once mutated can cause cancer. In fact, it appears to be a master switch for cancer. The primary name recommended by OMIM (The Online Mendelian Inheritance in Man database) was "Zinc Finger and BTB domain-containing protein 7".

In plant science databases the Maize Genome Database (GDB) standardizes nomenclature and symbols as follows: "the name and symbols that have been used for maize gene should be retained. The name and symbols of a gene should be represented with lower-case, italic character. Genes must be given 3 letter symbols. Newly detected maize genes that have been previously identified in other plant species should be named where appropriate with reference to the list of genetic names compiled by the commission on plant gene nomenclature. Symbols may describe a mutant phenotype or some aspect of gene structure or function".
But also here ambiguity of plant genes with English words has been reported, leading to distinctly bizarre cases such as in model plant *Arabidopsis* (genus of the family *Brassicaceae*): "*superman*" and "*clark kent*" mutants have extra stamens (male genitalia) in their flowers; the "*kryptonite*" mutation suppresses the function of "*superman*"; "*werewolf*", were plants have exceptionally hairy roots, and "*antikevorkian*" in which the programmed death of three out of the four female meiotic products is prevented in this mutant (dr. Kevorkian was the infamous American physician helping people to commit suicide).

This ambiguity is in part due to the guidelines of genes and protein nomenclature for the corresponding model organism as shown before. It is evident that a descriptive and free nomenclature as it is used for *Drosophila* makes automated identification of gene names very difficult, while a stringent nomenclature as it is used for yeast allows an easier identification of gene names.

Another problem that occurs is the tendency for error propagation with names based upon sequence similarity alone. For example, a gene named YFG2 is based upon sequence similarity to YFG1; gene YFG3 is then named based on similarity to YFG2 and YFG4 is named based upon similarity to YFG3. In fact, YFG3 and YFG4 may be quite distantly related to YFG1 so that in this case the relationship inferred by its name is misleading.

1.7 Terminology: principal link between the literature and ontology

The principal link between text and ontology is a terminology, which aims to map concepts to terms, but term variation and ambiguity make the integration of information available in text and ontology difficult. Term variation originates from the ability of a natural language to express a single concept in a number of ways as we have seen from the aforementioned examples. Term ambiguity occurs when the same term is used to refer to multiple concepts and is inherent to the biological and biomedical field as the evolution of species gave rise to many homologues and analogues.
Furthermore, biomedical and biological terms often appear in abbreviated forms. Although several methods have been develop to capture the different acronyms formed in the literature (Rimer and O'Connell 1998; Frantzi, S. Ananiadou et al. 2000; Rindflesch, Tanabe et al. 2000; Yoshida, Fukuda et al. 2000; Chang, Schutze et al. 2002;

Yu, Hatzivassiloglou et al. 2002; Yu, Hripcsak et al. 2002; Hisamitsu and J. Tsujii 2003; Schwartz and Hearst 2003; Adar 2004), there is still a need to assist the biologist in choosing the right term (acronym) that unambiguously pertain to a concept of interest (biological function).

Applications such as manual literature search, automated text-mining, name entity identification, gene or protein annotation, and linking of knowledge from different information source all require the knowledge of all used names referring to a given gene or protein (Fundel and Zimmer 2006). It is thus desirable to encourage researchers to use well-formed and approved gene and protein names that comply with strict nomenclature rules, easing literature search and automatic text processing.

An example of such an attempt is the UniProt consortium which is concerned with integrating information in the UniProt Knowledge Base. This consortium aims to provide a central, stable, comprehensive, fully classified, and richly and accurately annotated protein sequence database with extensive cross-references to other data sources. The expectation of the UniProt project is that the SwissProt/UniProt and Entrez genes will increasingly share their nomenclature and that the mapping between databases will be increasingly complete and unambiguous. This will facilitate the generation of gene name dictionaries for text mining application.

Nevertheless, this will not remove the difficulty entirely, since there are still huge numbers of published documents around containing "legacy" and add-hoc terms that need to be integrated and analysed. Therefore it is paramount to develop systems that can resolve ambiguity problems in a general way and irrespectively of the organism under consideration. This will be the scope of ***chapters 4 and 5***.

1.8 Text mining techniques

Text mining directed towards knowledge discovery comprises various steps or techniques, e.g. Entity Recognition (ER), Information Retrieval (IR), Information Extraction (IE) and Data Mining (DM). It is more strictly defined as the "discovery by computer of new, previously unknown information, by automatically extracting information from different written resources". Therefore, IE does not qualify as a text-mining tool itself, as it can only extract what has already been published and thus rather forms the basis for text mining in the same way that ER forms the basis for IE.

1.8.1 ER: Entity Recognition

The objective of entity recognition (ER) is to find the biological entities that are mentioned within a text, in particular the names of genes and proteins. At first glance ER might seem neither challenging nor particularly useful, but in fact it is probably the most challenging task in biomedical text mining and a prerequisite for both IE and IR. Basic approaches to find named entities include rule-based techniques using finite-state transducers (Roche E. and Schabes Y. 1997; Cunningham H., Maynard D. et al. 2000) and statistical taggers that use Support Vector Machines (SVMs) (Li Y., Bontcheva K. et al. 2005) or Hidden Markov Models (HMMs) (Manning C.D. and Schutze 1999). Once detected, biomedical and biological terms need to be normalized (refer all term variants to a single descriptor) and grounded (link through identifier to entry in database e.g. UniProt). This illustrates how term variation and ambiguity can hamper the recognition of biological and biomedical entities, not only in the nomenclature using long terms but

also for their respective short forms. ER can also be usefully applied on its own in cross-linking literature that is related to certain genes as it is used in the Information Hyperlinked Over Protein system (iHOP) (Hoffmann and Valencia 2005).
A particularly common term variation type in biology is the representation by acronyms. In MEDLINE abstracts, 64,242 new acronyms were introduces in 2004, with an estimated total of 800,000 (Chang and Schutze 2006). Acronym recognition aims to extract pairs of short forms (acronyms) and their associated long form (expanded). State-of-the-art acronym recognition can be categorized in heuristics scoring rules (Schwartz and Hearst 2003; Adar 2004), machine learning (Pakhomov 2002) and statistical methods (Okazaki and Ananiadou 2006).

1.8.2 IR: Information Retrieval

Information retrieval (IR) is the activity of finding documents that answer an information need with the aid of indexes. One of the best-known and used IR systems in the wider public is probably Google. However, one of the drawbacks of such system is that the user is faced with reading many documents in order to discover the facts reported in them. There is a multitude of systems based on IR techniques and applied to databases of biological and biomedical literature with reasonable high precision including Textpresso (Muller, Kenny et al. 2004), iHOP (Hoffmann and Valencia 2005), GoPubMed (Doms and Schroeder 2005), EBIMed (Rebholz-Schuhmann, Kirsch et al. 2007), PubMatrix (Becker, Hosack et al. 2003), PubFinder (Goetz and von der Lieth 2005) , MedScan, LitMiner, Chilibot, Transminer and BioRAT.
The best-known biomedical IR system, PubMed is an *ad hoc* system that uses two established IR methodologies, the Boolean model and the vector model. The Boolean model enables the user to retrieve all documents containing certain combinations of terms by using a logical operation, for example "Alk1 AND damage response". In contrast, the vector model represents each document by a term vector, in which each term is assigned a value according to a frequency-based weighting scheme. These document vectors can subsequently be compared to a query vector that specifies the relative importance of each query term. Alternatively they can be compared to each other to calculate document similarity, which is used in PubMed by the "related articles" function. However, it is crucial not to restrict IR to exact matching of query terms, because term ambiguity and variation phenomena may cause irrelevant information to be retrieved ("low precision") and relevant information to be missed or overlooked ("low recall"). (van Driel et al, Eur J Hum Genet 2006)

1.8.3 IE: Information extraction

In contrast to IR, information extraction (IE) strives to extract information from texts without requiring the end user of the information to read the entire text. IE depends on named entity recognition (NER) as the main step in accessing textually described domain-specific information. IE can be used to support a fact-retrieval service or as a step towards text mining based on conceptually annotated text. Furthermore, IE can be ontology-based, i.e. map a term occurring in a text to a concept in ontology, typically in the absence of any explicit link between term and concept. This is a passive ontology use. On the other hand, IE can be also ontology-driven, which means that it makes active use of ontology in processing in order to strongly guide and constrain analysis.
Two main examples for approaches that extract relationships from biological texts are co-occurrence methods and natural language processing (NLP). Co-occurring terms in a discourse assumes a mutual relationship. NLP strive to keep the semantic of the

relationship of terms under investigation in a discourse. Co-occurrences methods tend to give a better recall but a weaker precision as compared to NLP methods, and thus are well suited as part of exploratory tools because of their ability to identify relationship of almost any type (Jensen, Saric et al. 2006). Borrowing terminology from logic, precision may be viewed as the "degree of soundness" and recall as the "degree of completeness". Text mining using natural language processing (NLP) not only uses sentence structure, but employs part of speech (POS) and phrase recognition to identify certain relationship among entities in a sentence (Hunter and Cohen 2006) . However, due to the inherent complexities of retrieving a sound meaning from a group of sentences that use complex grammar, NLP methods in text mining are often unreliable in identifying a relationship between multiple sentences (Ding, Berleant et al. 2002; Daraselia, Yuryev et al. 2004), also known as anaphoric relationship. An anaphoric relationship is exemplified by the following two sentences: "The dog was sick. It had to be put down". Moreover, complex sentences that contain multiple relationships give rise to additional, erroneous relationships.
Furthermore, co-occurrence is unable to extract direct and indirect relationships, for instance whether or not a compound X directly phosphorylates a compound Y, whereas NLP combines the analysis of syntax and semantics and can therefore – in principle - tackle the issue of direct and indirect relationships.

1.9 Text mining legacy (literature-based discovery and hypothesis generation)

The following section will focus on the challenges of terminology and terminological processing and novel techniques for information extraction in text mining. Text mining can be used for a multitude of purposes, for instance to interpret gene expression clustering or to model complex biological pathways based on published literature (Blaschke, Oliveros et al. 2001; Glenisson, Coessens et al. 2004).
More importantly, text mining also serves the purpose of hypothesis generation and biological discovery. For example, one set of literature has shown that dietary fish oils lead to certain vascular changes, and a separate set of literature has reported that such vascular changes would benefit patients with Raynaud's syndrome. Raynaud's syndrome is a condition that affects blood flow to the extremities such as fingers, toes, nose and ears, when exposed to temperature changes or stress. In a similar fashion text mining was also applied to infer that migraines may be caused by magnesium deficiency and that the connection between arginine intake and blood level of somatomedins may be critical in the treatment of thymic deficiencies.
Besides finding relationships between diseases and potential therapeutic interventions in the literature, enriching protein-protein interactions and extracting scientific abstracts pertaining to a topic of interest, text mining has been successfully applied to much more.
Text mining can aid in finding new trends in different research field as shown in the example of Rebholz-Schuhmann and colleagues (Rebholz-Schuhmann, Cameron et al. 2007). Upon systematic analysis of the scientific literature from medical informatics and bioinformatics research they concluded that emerging topics, equally important to bioinformatics and medical informatics in recent years are: microarray experimentation, ontologies, open source, text mining, and support vector machines. Emerging topics that evolved only in bioinformatics were systems biology, protein interaction networks and statistical methods for microarray analysis, whereas emerging topic in medical informatics were grid technology and tissue microarrays. Thus both fields share a common technological development that tends to be initiated by new developments in biotechnology and computer science. Another example for predicting new research aspects that are about to become popular is the research of future "hot" proteins that

can become commercially attractive targets for the development of antibodies and inhibitors.

Furthermore TM can be used to search for correlated events as exemplified by the Amazon's, Ebay's and other e-businesses' function "Customers who bought this item also bought ...". This can be similarly applied in life sciences to discover fundamental properties of regulatory networks and uncovered relationships between biological entities for example. Text mining has even been shown to enhance microarray gene expression analysis by incorporating biological information mined from the literature into standard distance metric-based clustering algorithms to construct more accurate gene relation networks than the standard clustering algorithms alone (Karopka, Scheel et al. 2004).
Text mining as a tool proved even more valuable when combined with other data types. Integration of literature-based protein networks and studies of linkage mapping identified candidate genes for Alzheimer diseases within a genomic region on the basis of their interactions with genes that are already known to cause the disease (Krauthammer, Kaufmann et al. 2004).

As such text mining in life sciences should be assessed according to its contribution to biology-driven problems to maintain the momentum gained over the past decades. After this illustration of text mining's assets, we come to the important question of how to select for the best text mining tools. The effectiveness of text mining tools is usually reported using certain metrics that will be explained in the subsequent section.

1.10 Text mining evaluation

For classification problems, bioinformaticians usually measure the performance of a model in terms of error rate: the percentage of incorrectly classified instances in the data set. Usually a model is build in order to classify new data and thus the performance of a model on "unseen" data is of interest. The training set (seen data) is used to build the model, i.e. to determine its parameters, and the test set (unseen data) to measure its performance, i.e. holding the parameters constant. Sometimes a validation set is required to tune the model, for example for pruning a decision tree. However, the validation set cannot be used for testing, as it does not qualify as unseen data.
Training data, test data and validation data have to be representative samples of the data that the model will be applied to. If a lot of data are available, two independent samples are selected, one used for training and one for testing. The more training data are available, the higher the quality of the model and the more test data, the more accurate the error estimate.
Major drawbacks in obtaining big data sets are their expense and time consumption, therefore a limited data set is usually selected and a holdout procedure is applied. In other words, a random split is done of the data to generate a test set and a training data. Typically 1/3 and 1/10 are held out for testing. However, the split into training and test data risks to be non-representative, i.e. a certain class is not represented in the training set and thus the model will not learn to classify it. In such condition a stratified holdout is applied, i.e. the data is sampled in such a way that each class is represented in both sets. Unfortunately this procedure does not work well on smaller data sets, which require a maximisation of data utilization.
One solution to the latter is a k-fold crossvalidation that divides the data randomly into k subsets of equal size. The model is trained on k-1 subsets and one subset is used for testing. This process is repeated k times (folds) so that all subsets are used exactly once for testing. Finally the average performance is computed on the k test sets. K-fold

crossvalidation effectively uses all the data for both training and testing. Typically $k=10$ is used. Another variant of k-fold crossvalidation is the leave-one-out crossvalidation. Leave-one-out crossvalidation is simply k-fold crossvalidation with k set to n, the number of instances in the data set. This means that the set consists only of a single instance, which will be classified either correctly or incorrectly. This technique thus maximises the use of data, i.e. the training is done on n-1 instances. In contrast, leave-one-out is not suitable for large data sets because of the high computational cost of the required large number of training runs.

The reported performance of a classified data set can be summarized in a confusion matrix or contingency table as shown in Table 2. Assuming a two-way classification, four classification outcomes are possible such as displayed in our example contingency table. Here True Positives (TP) are class members correctly classified as class members; True Negatives (TN) are class non-members classified as non-members; False Positives (FP) are class non-members incorrectly classified as class members, and False Negatives (FN) are class members classified as class non-members.

Table 2: Confusion matrix or contingency table

Category set $C=\{c_1,\ldots,c_{\|C\|}\}$		Expert Judgment	
		YES	NO
Classifier judgement	YES	$TP=\sum_{i=1}^{\|C\|}TP_i$	$FP=\sum_{i=1}^{\|C\|}FP_i$
	NO	$FN=\sum_{i=1}^{\|C\|}FN_i$	$TN=\sum_{i=1}^{\|C\|}TN_i$

Measures that are commonly used in information retrieval, classifications tasks and text mining are precision and recall. Precision measures the number of class members classified correctly over the total number of instances classified as class members. The recall reports the number of class members classified correctly over the total number of class members. Precision and recall can be combined into the F-measure, which is simply the harmonic mean of precision and recall. The F-measure is used if both precision and recall are equally important.

Independently of the way the performance of the model is measured, the performance measure is always carried out in unseen data sets, i.e. test set, but never on seen data, i.e. the training set. Performance on the training set only tells us that the model learned what it was supposed to learn, hence is not a good indicator of the performance on the unseen data.

As defined here, precision and recall are to be understood as subjective probabilities, which mean they measure the expectation of the user that the system will behave correctly when classifying an unseen document under a given category.

Two different methods may be adopted for obtaining estimates of precision and recall:

-Microaveraging: precision and recall are obtained by summing over all individual decisions, therefore summing over category-specific contingency table to generate the "global" contingency table.

$$\text{Precision}=\frac{TP}{TP+FP}=\frac{\sum_{i=1}^{|c|}TP_i}{\sum_{i=1}^{|c|}(TP_i+FP_i)},\ \text{Recall}=\frac{TP}{TP+FN}=\frac{\sum_{i=1}^{|c|}TP_i}{\sum_{i=1}^{|c|}(TP_i+FN_i)},\ \text{and}$$

$$F - measure = \frac{2 * (\Pr ecision * \Re call)}{\Pr ecision + \Re call}$$

-Macroaveraging: Precision and recall are evaluated first "locally" for each category and then "globally" by averaging over the results of the different categories.

$$\Pr ecision = \frac{\sum_{i=1}^{|C|} precision_i}{|C|}, \qquad \Re call = \frac{\sum_{i=1}^{|C|} recall_i}{|C|}, \qquad \text{and}$$

$$F - measure = \frac{2 * (\Pr ecision * \Re call)}{\Pr ecision + \Re call}$$

These two methods may give quite different results, especially if the various categories show a very different generality. For example, the ability of a classifier to behave well also on categories with low generality (category with few positive training instances) will be emphasized by macroaveraging and much less by microaveraging.
Similarly the ***kappa*** coefficient statistics (Cohen 1960; Carletta 1996) can be derived from the above contingency table. Kappa statistics tests the null hypothesis that there is no more agreement than might occur by chance given a random guessing. Kappa ranges from 0 for chance agreement to 1 for full agreement. However, with the kappa statistics, the agreement between different annotators (inter-annotator agreement) is measured which contrasts to the case of precision and recall mentioned above where the agreement is between an annotator or expert and a computer or machine. Thus, the inter-annotator agreement allows conclusion about the stability of annotation, while the agreement for each annotator with himself (intra-annotator agreement) indicates the reproducibility of the annotation (Gut and P. Bayer 2004).

1.11 Outlook on this thesis

As can be taken from the information given in the above sections of the introduction, text mining is a gaining importance in the life sciences and thus the main scope of this thesis. Addressing the entire set of problems bioinformatics faces in this genomic era is virtually impossible in the frame of just one PhD thesis. Therefore we will address a subset of these tasks here in order to improve our understanding of biological phenomena. We will adopt an integrative algorithmic approach to navigate from sequence analysis to understanding unstructured texts, i.e. scientific literature.
In this introduction (***Chapter 1***) we attempted to highlight various aspects - without a claim of exhaustiveness - of the main benefits of Bioinformatics in easing our understanding of some biological phenomena at the molecular level as well as some challenges Bioinformatician addresses. ***Chapter 2*** adopts an integrative approach to concatenate protein sequence domain prediction from different prediction methods to support relatedness of proteins under investigation.
Chapter 3 endeavours to mine Medline based on term co-occurrence and faces the issue of gene nomenclature ambiguity, whereas ***Chapter 4*** addresses the gene nomenclature ambiguity in all species and proposes an approach of efficient resolution of the ambiguity. ***Chapter 5*** provides an application of the methodology described in ***Chapter 4***, namely an effective and unambiguous query formulation for literature retrieval. Finally this thesis concludes with ***Chapter 6*** summarising the main findings and the

contribution of this work to scientific research, as well as some propositions for future directions of this work.

References

(2006). "The Gene Ontology (GO) project in 2006." Nucleic Acids Res **34**(Database issue): D322-6.

Adar, E. (2004). "SaRAD: a Simple and Robust Abbreviation Dictionary." Bioinformatics **20**(4): 527-33.

Alako, B. T., A. Veldhoven, et al. (2005). "CoPub Mapper: mining MEDLINE based on search term co-publication." BMC Bioinformatics **6**: 51.

Altschul, S. F., T. L. Madden, et al. (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Res **25**(17): 3389-402.

Baldi P and B. S. (2001). "Bioinformatics. The Machine Learning Approach." MIT Press.

Becker, K. G., D. A. Hosack, et al. (2003). "PubMatrix: a tool for multiplex literature mining." BMC Bioinformatics **4**: 61.

Blaschke, C., J. C. Oliveros, et al. (2001). "Mining functional information associated with expression arrays." Funct Integr Genomics **1**(4): 256-68.

Bodenreider, O. (2004). "The Unified Medical Language System (UMLS): integrating biomedical terminology." Nucleic Acids Res **32**(Database issue): D267-70.

Carletta, J. C. (1996). "Assessing Agreement on Classification Tasks: The Kappa Statistic." Computational Linguistics **Vol. 22, No 2**(2): 249-254.

Chang, J. T. and H. Schutze (2006). Abbreviations in biomedical text. In Text Mining for Biology and Biomedicine (Ananiadou, S. McNaught, J. eds), ARTECH House: 138-165.

Chang, J. T., H. Schutze, et al. (2002). "Creating an online dictionary of abbreviations from MEDLINE." J Am Med Inform Assoc **9**(6): 612-20.

Cohen, J. (1960). "A Coefficient of Agreement for Nominal Scales." Educational and Psychological Measurement **Vol. 20**(No 1.): pp. 37-46.

Cunningham H., Maynard D., et al. (2000). "JAPE: a Java Annotation Patterns Engine (Second Edition)." Technical report, University of Sheffield, Department of Computer Science.

Daraselia, N., A. Yuryev, et al. (2004). "Extracting human protein interactions from MEDLINE using a full-sentence parser." Bioinformatics **20**(5): 604-11.

Ding, J., D. Berleant, et al. (2002). "Mining MEDLINE: abstracts, sentences, or phrases?" Pac Symp Biocomput: 326-37.

Doms, A. and M. Schroeder (2005). "GoPubMed: exploring PubMed with the Gene Ontology." Nucleic Acids Res **33**(Web Server issue): W783-6.

Edgar, R. C. (2004). "MUSCLE: multiple sequence alignment with high accuracy and high throughput." Nucleic Acids Res **32**(5): 1792-7.

Frantzi, K., S. Ananiadou, et al. (2000). "Automatic recognition of multi-word Terms: The C-value/NC-value method." International Journal on Digital Libraries **3**(2): 115-130.

Fundel, K. and R. Zimmer (2006). "Gene and protein nomenclature in public databases." BMC Bioinformatics **7**: 372.

Glenisson, P., B. Coessens, et al. (2004). "TXTGate: profiling gene groups with text-based information." Genome Biol **5**(6): R43.

Goetz, T. and C. W. von der Lieth (2005). "PubFinder: a tool for improving retrieval rate of relevant PubMed abstracts." Nucleic Acids Res **33**(Web Server issue): W774-8.

Gut, U. and P. Bayer (2004). "Measuring the Reliability of Manual Annotations of Speech Corpora." Proc. Int. Conf. on speech Prosody (SP): 565-568.

Hamosh, A., A. F. Scott, et al. (2005). "Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders." Nucleic Acids Res **33**(Database issue): D514-7.

Hisamitsu, T. and J. Tsujii (2003). Measuring Term Representiveness. Information Extraction in the Web Era. Measuring Term Representiveness. Information Extraction in the Web Era, LNAI 2700, Springer: 45-76.

Hoffmann, R. and A. Valencia (2005). "Implementing the iHOP concept for navigation of biomedical literature." Bioinformatics **21 Suppl 2**: ii252-8.

Hunter, L. and K. B. Cohen (2006). "Biomedical language processing: what's beyond PubMed?" Mol Cell **21**(5): 589-94.

van Driel MA, Bruggeman J, Vriend G, Brunner HG, Leunissen JAM (2006) "A text mining analysis of the human phenome." Eur. J. Human Genet. 14, 535-542.

Jensen, L. J., J. Saric, et al. (2006). "Literature mining for the biologist: from information retrieval to biological discovery." Nat Rev Genet **7**(2): 119-29.

Karopka, T., T. Scheel, et al. (2004). "Automatic construction of gene relation networks using text mining and gene expression data." Med Inform Internet Med **29**(2): 169-83.

Krallinger, M., R. A. Erhardt, et al. (2005). "Text-mining approaches in molecular biology and biomedicine." Drug Discov Today **10**(6): 439-45.

Krauthammer, M., C. A. Kaufmann, et al. (2004). "Molecular triangulation: bridging linkage and molecular-network information for identifying candidate genes in Alzheimer's disease." Proc Natl Acad Sci U S A **101**(42): 15148-53.

Li Y., Bontcheva K., et al. (2005). "Using Uneven Margins SVM and Perceptron for Information Extraction." In Proceedings of Ninth Conference on Computational Natural Language Learning (CoNLL).

Liu, F., T. K. Jenssen, et al. (2004). "FigSearch: a figure legend indexing and classification system." Bioinformatics **20**(16): 2880-2.

Ma, L., B. G, et al. (2007). "ClustalW and ClustalX version 2.0." Bioinformatics.

Manning C.D. and H. Schutze (1999). "Foundations of Statistical Natural Language Processing." The MIT Press.

Muller, H. M., E. E. Kenny, et al. (2004). "Textpresso: an ontology-based information retrieval and extraction system for biological literature." PLoS Biol **2**(11): e309.

Notredame, C., D. G. Higgins, et al. (2000). "T-Coffee: A novel method for fast and accurate multiple sequence alignment." J Mol Biol **302**(1): 205-17.

Okazaki, N. and S. Ananiadou (2006). A term recognition approach to acronym recognition. In Proceedings of Coling/ACL Conference 2006.

Pakhomov, S. (2002). "Semi-supervised maximum entropy-based approach to acronym and abbreviation normalization in medical texts." In Association for Computational Linguistics (ACL): 160-167.

Rebholz-Schuhmann, D., G. Cameron, et al. (2007). "SYMBIOmatics: synergies in Medical Informatics and Bioinformatics--exploring current scientific literature for emerging topics." BMC Bioinformatics **8 Suppl 1**: S18.

Rebholz-Schuhmann, D., H. Kirsch, et al. (2007). "EBIMed--text crunching to gather facts for proteins from Medline." Bioinformatics **23**(2): e237-44.

Rimer, M. and M. O'Connell (1998). "BioABACUS: a database of abbreviations and acronyms in biotechnology and computer science." Bioinformatics **14**(10): 888-9.

Rindflesch, T. C., L. Tanabe, et al. (2000). "EDGAR: extraction of drugs, genes and relations from the biomedical literature." Pac Symp Biocomput: 517-28.

Roche E. and Schabes Y. (1997). " Finite-State Language Processing." MIT Press.

Ruepp, A., O. N. Doudieu, et al. (2006). "The Mouse Functional Genome Database (MfunGD): functional annotation of proteins in the light of their cellular context." Nucleic Acids Res **34**(Database issue): D568-71.

Schwartz, A. S. and M. A. Hearst (2003). "A simple algorithm for identifying abbreviation definitions in biomedical text." Pac Symp Biocomput: 451-62.

Sherlock, G. (2001). "Analysis of large-scale gene expression data." Brief Bioinform **2**(4): 350-62.

Smalheiser, N. R. and D. R. Swanson (1998). "Using ARROWSMITH: a computer-assisted approach to formulating and assessing scientific hypotheses." Comput Methods Programs Biomed **57**(3): 149-53.

Spasic, I., S. Ananiadou, et al. (2005). "Text mining and ontologies in biomedicine: making sense of raw text." Brief Bioinform **6**(3): 239-51.

Spellman, P. T. and et al. (1988). "Comprehensive Identification of Cell Cycle regulated Genes of the Yeast Saccharomyces Cerevisiae by Microarray Hybridization." Mol. Biol. Cell **Vol. 9**: 3273-3297.

Tamayo, P. and et al. (1988). "Interpreting patterns of Genes Expression with Self-Organizing Maps: Methods and Application to Hematopoietic Differenciation." **Vol. 96**: 2907-2912.

Tanabe, L., U. Scherf, et al. (1999). "MedMiner: an Internet text-mining tool for biomedical information, with application to gene expression profiling." Biotechniques **27**(6): 1210-4, 1216-7.

Tatusov, R. L., N. D. Fedorova, et al. (2003). "The COG database: an updated version includes eukaryotes." BMC Bioinformatics **4**: 41.

Tu, Q., H. Tang, et al. (2004). "MedBlast: searching articles related to a biological sequence." Bioinformatics **20**(1): 75-7.

Tuason, O., L. Chen, et al. (2004). "Biological nomenclatures: a source of lexical knowledge and ambiguity." Pac Symp Biocomput: 238-49.

Weeber, M., B. J. Schijvenaars, et al. (2003). "Ambiguity of human gene symbols in LocusLink and MEDLINE: creating an inventory and a disambiguation test collection." AMIA Annu Symp Proc: 704-8.

Wheeler, D. L., T. Barrett, et al. (2006). "Database resources of the National Center for Biotechnology Information." Nucleic Acids Res **34**(Database issue): D173-80.

Yoshida, M., K. Fukuda, et al. (2000). "PNAD-CSS: a workbench for constructing a protein name abbreviation dictionary." Bioinformatics **16**(2): 169-75.

Yu, H., V. Hatzivassiloglou, et al. (2002). "Automatically identifying gene/protein terms in MEDLINE abstracts." J Biomed Inform **35**(5-6): 322-30.

Yu, H., G. Hripcsak, et al. (2002). "Mapping abbreviations to full forms in biomedical articles." J Am Med Inform Assoc **9**(3): 262-72.

Zhou, G., X. Wen, et al. (2004). "B.E.A.R. GeneInfo: a tool for identifying gene-related biomedical publications through user modifiable queries." BMC Bioinformatics **5**: 46.

Chapter 2

TreeDomViewer, a tool for the visualisation of phylogeny and protein domain structure

Blaise T.F. Alako [1,2], Daphne Rainey [3], Harm Nijveen [1], and Jack A.M. Leunissen [1*]

1. Laboratory of Bioinformatics, Wageningen University and Research Centre, PO Box 8128, 6700 ET Wageningen, The Netherlands
2. Centre for BioSystems Genomics, PO Box 98, 6700 AB Wageningen, The Netherlands
3. KEYGENE NV, PO Box 216 6700 AE Wageningen, The Netherlands

Nucleic Acids Res (2006), 34, W104-109.
(With permission)

Abstract

Phylogenetic analysis and examination of protein domains allow accurate genome annotation and are invaluable to study proteins and protein complex evolution. However, two sequences can be homologous without sharing statistically significant amino acid or nucleotide identity, presenting a challenging bioinformatics problem.
We present TreeDomViewer, a visualization tool available as a web-based interface that combines phylogenetic tree description, multiple sequence alignment and InterProScan data of sequences and generates a phylogenetic tree projecting the corresponding protein domain information onto the multiple sequence alignment. Thereby it makes use of existing domain prediction tools such as InterProScan. TreeDomViewer adopts an evolutionary perspective on how domain structure of two or more sequences can be aligned and compared, to subsequently infer the function of an unknown homolog. This provides insight into the function assignment of, in terms of amino acid substitution very divergent but yet closely related family members. Our tool produces an interactive scalar vector graphic (SVG) image that provides orthological relationship and domain content of proteins of interest at one glance. Alternatively PDF, JPEG or PNG formatted output can also be provided.
These features make TreeDomViewer a valuable addition to the annotation pipeline of unknown gene or gene product. TreeDomViewer is available at
http://www.bioinformatics.nl/tools/treedom/.

Introduction

The past years have seen the rapid sequencing of genomes from many different organisms. Sequencing itself is no longer the bottleneck in genomic studies; the bottleneck is reliable annotation of new genes. Information from widely studied model species included in comparative annotation genomics has greatly aided in these annotation efforts and proofed to be a powerful tool. Quoting Constantinesco et al. (1): "comparative genomic studies have been invaluable to the annotation efforts in addition to their contribution to the understanding of protein evolution". Sometimes homologous gene products have strong sequence similarities so that the inference of homology is straightforward. However, accumulation of multiple substitutions in the course of divergent evolution can make homologous sequences as dissimilar as any two proteins chosen randomly from a database (2).

Several bioinformatics approaches have been developed to identify remote homology in the absence of pairwise similarity, one of the popular ones being protein fold recognition (FR) (3). Briefly, FR detects homology based on a combination of evolutionary criteria and structural considerations. FR differs from traditional sequence homology database searches insofar as the databases to be searched by FR contain only proteins with experimentally determined structure rather than all protein sequences. Hence the availability of a related structure in the Protein Data Bank is an essential but not sufficient prerequisite for the success of FR-based identification of homologs (4). However, homology is defined on the basis of evolution rather than function. Homologues can fulfil different functions and share only very general similarities; even orthologs may fulfil non-identical roles states (5).

Moreover, homology is not necessarily a one-to-one relationship, because a single gene in one genome may correspond to a whole family of paralogs in another genome, which may be functionally diverse. Hence there is a pitfall of over-prediction (i.e. too specific functional assignment) to be avoided when annotating unknown protein or gene function by homology, using either simple or sophisticated existing bioinformatics tools (4).

Currently there is a multitude of tools available for the visualization of information contained within a protein sequence such as signal peptides (6), transmembrane domains (7,8) and functional domains such as InterProScan (9). The latter currently comprises fifteen domain prediction methods.

However, until now there is no tool available combining in one view protein sequence analysis with orthological information, thereby essentially combining proteomics information with phylogenomics (see e.g. (10)) independent of the available 3D structure in databases.

In this paper we present a more convenient way of identifying putative family members based on their evolutionary history and supported by their conserved structural domains, as the evolution of the later, unlike amino sequence substitutions, occurs at a slower rate throughout evolution.

This phylogenetic visualization tool allows a rapid 'first pass' quality screening of search results from InterProScan and others (e.g. the EMBOSS package (11)). One of its strengths is the forthright generation of a publication-quality graphic output. TreeDomViewer is available as a Perl-based web interface that accepts a multiple sequence file in any common format as input and produces a phylogenetic tree with the corresponding protein domain information projected onto the multiple sequence alignment next to it. Although a powerful tool by itself, TreeDomViewer is obviously dependent on the quality of the analysis tools and multiple alignments.

Implementation and Design

Data preparation and processing

The minimal input required by TreeDomViewer is a set of aligned or unaligned sequences. In case where the input file is a sequence file solely, ClustalW (12) is used to align the sequences and a tree description is calculated subsequently using ClustalW's built-in neighbor-joining option (13).
By default TreeDomViewer combines the output from several programs, i.e. a multiple alignment (in any common sequence format, such as FASTA or Clustal), a phylogenetic tree (in standard Newick or PHYLIP format (14)) and domain predictions (in InterProScan's "raw" format).
The ability to upload precalculated files makes the tool extremely flexible, as the user may want to edit manually his/her multiple sequence alignment or other input files, or select another program for alignment of phylogeny construction than the ones provided by TreeDomViewer.
There are two possibilities to run TreeDomViewer, either interactive, where the user uploads the sequence and/or (alignment, tree description file and the InterProScan analysis file), or in batch mode: the user uploads either the sequence or multiple alignment file but not the InterProScan file. He/she will receive links to the result via email upon job completion and get the option of saving input files as this will save time for future runs of the same data set. The tool is sufficiently sophisticated to decide which prediction method is the most time consuming one and if selected it may automatically switch to batch mode.
The tool combines multiple domains on the same picture and it is necessary to have them sorted by domain length in order to have the largest domain drawn first. This provides a quick overview of multiple predictions on the same region.
One feature of major importance in TreeDomViewer is the alignment of structural domains. This allows for quick checking of the alignment quality, easy inference of homology even when the sequence residue similarity is very low, and support of the phylogeny based on functional characteristics evidences.
The rate-limiting step in TreeDomViewer is the computation of the structural domains using InterProScan. By running these calculations in parallel on 10 nodes of a small Linux cluster, turn-around times are still acceptable. For example, the analysis of 60 protein sequences of 1000 amino residues each is performed in less than 3 minutes.

Design overview

TreeDomViewer is implemented in Perl as a web based service running on an Apache 2.0 web server on a Linux platform (SuSE linux Enterprise Server 9). The core application consists of three main programs: *Sygtree, Treedom and Clustalw*. The first two programs are full command line tools written in-house in C and Perl respectively and can be used as plug-in for other applications. A web interface was built on top of these programs via a Perl CGI script (figure 1).

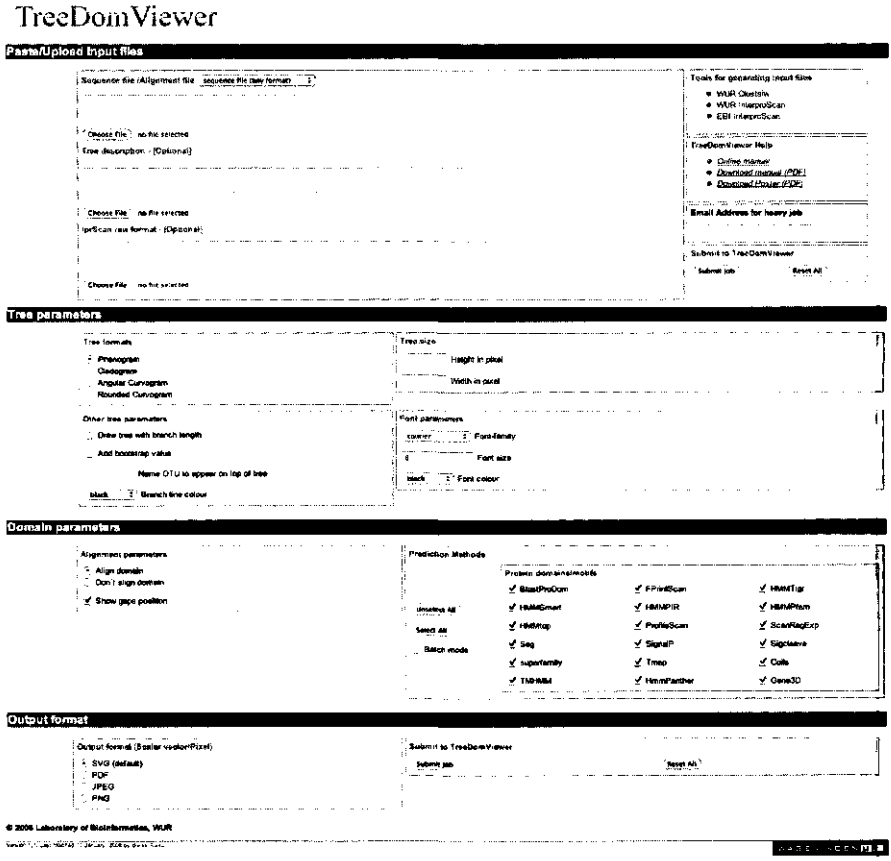


Figure 1: TreeDomViewer web-based interface. Alternative means of generating the input file are provided on the top-right panel.

This preserves platform independence across multiple operating systems and allows the user to interact with the different *TreeDomViewer* programs without computer programming or (shell) scripting skills. A global overview of TreeDomViewer workflow is presented in figure 2. Full explanation of the tool's mode of action is available as an online or downloadable (PDF) manual at the web-interface.

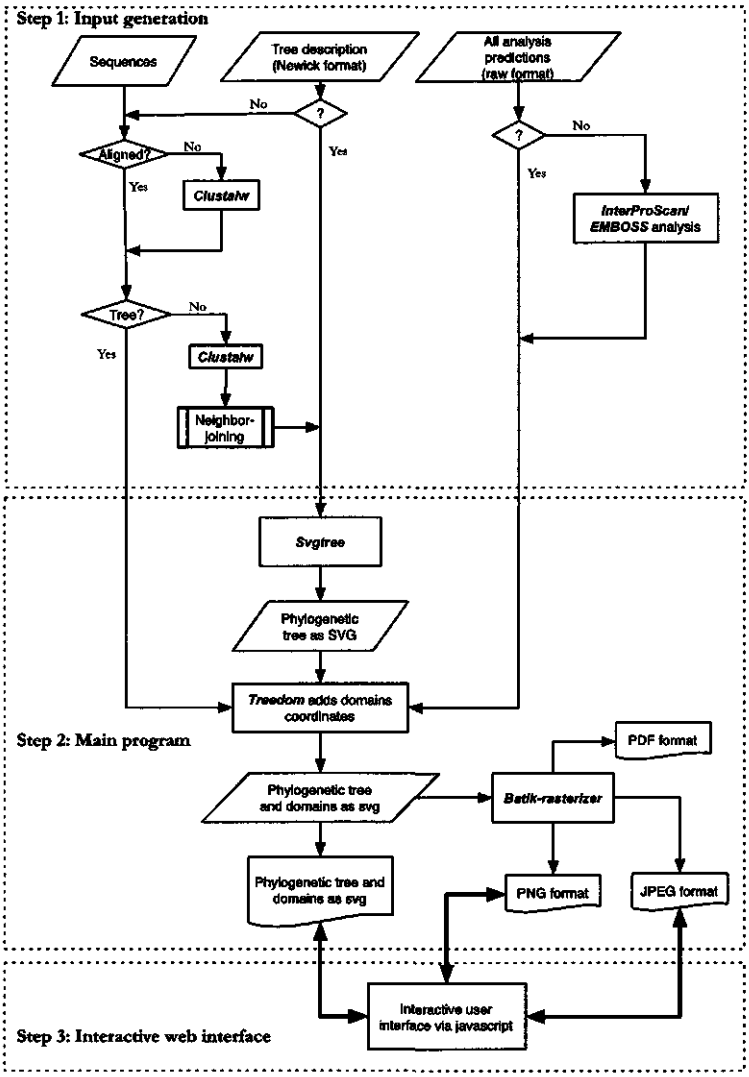


Figure 2: Flowchart of TreeDomViewer illustrating sequence of application implemented. Software tools used are in bold. Three types of data input are processed and domain information is coordinated with the alignment and phylogenetic tree information to produce an interactive SVG output.

The software was developed on a Linux (SuSE 8.2 and SuSE linux Enterprise Server 9) platform and most of its modules were written from scratch to prevent dependency issues when migrating to newer versions of Linux OS or Perl.

The TreeDomViewer web interface was tested on Windows XP, Mac OS X and several flavors of Linux OS browsers with good results. Some JavaScript event handling problems for interacting with the SVG output were encountered on Mac OS X and Linux OS. This can be attributed to the web browsers used (konqueror, Mozilla, Opera), as at the moment no browser supports SVG to its full extend. Currently most browsers still require a SVG plug-in, downloadable from the Adobe site. However, the latest version of Mozilla Firefox browser (version 1.5) has already native (built-in) SVG support and it is to be expected that more browser will soon follow.

Most browsers handle SVG pictures quite well when standard shapes such as rectangles or lines were instructed to be drawn on the screen. In this matter TreeDomViewer takes it one step further by giving life to these shapes through JavaScript. As all browsers support and display JPEG (Joint Photographic Experts Group) and PNG (Portable Network Graphic), TreeDomViewer uses *batik-rasterizer* to provide them as an alternative output format besides PDF format, thereby circumventing the SVG plug-in flaw as noted above.
Batik-rasterizer is part of the open source Apache Batik toolkit 1.6 (http://xml.apache.org/batik/).
Most of SVG output features such as mouse-over events are retained except zoom-in and zoom-out. As we aimed at integrating as much information as possible within a single picture, domain predictions are linked to their source database where more information can be retrieved.

Output description

By default TreeDomViewer provides SVG output of the tree and domain information. The user's web browser needs to be SVG-enabled in order to view the output. Conveniently, the viewer first checks the web browser to clarify whether it is SVG-enabled or not, and fetches the Adobe SVG plug-in (http://www.adobe.com/svg/viewer/install/) and prompts for its installation if needed.
The user can change parameters for the tree plotting such as tree format, set to phenogram as default, and many more features as shown in figure 1. Links to individual protein analysis tools are also provided. It is noteworthy that TreeDomViewer does not execute protein analysis on its own, but instead provides an interface to InterProScan and other programs as shown in the prediction method section of its interface.
There are several interactive features such as zoom-in and zoom-out, mouse-over access for information on each domain, references to techniques used to produce the domain, and on-the-fly switching on and switching off of domain prediction through the left control panel (figure 3 as well as an accompanying legend of the graphic).

Alternative formats such as PDF, JPG and PNG are also provided. Although TreeDomViewer was designed for protein analysis, nucleotide sequences can be handled as well. Moreover, TreeDomViewer is able to generate the output of any domain prediction tool, making it the visualization tool of choice at any level of functional or phylogenetic study. Tools such as Adobe Illustrator can be used to manipulate domain colors of TreeDomViewer SVG file.
In order to illustrate our approach we analyzed a subset of the lipocalin family members. Lipocalins are a superfamily of proteins that carry hydrophobic prosthetic groups. Lipocalins share a very low sequence similarity, hence it can be expected to be a cumbersome affair to infer homology with the conventional sequence similarity or identity techniques. To further our illustration a subset of the lipocalins was selected manually in accordance with those reported by Ganfornina et al. (15). We chose this family to illustrate the features of TreeDomViewer because of their strong divergent protein sequence, denoting a rapid rate of molecular evolution, Moreover, the evolutionary history of the lipocalins is rich in gene duplication events, which increases the difficulty of obtaining an understanding of orthologous relationships. As denoted by red features in figure 3, there are three conserved sequence motifs called structurally conserved regions (SCRs) that have been proposed by Flower et al. (16) as a prerequisite for a protein to be considered as a lipocalin.

Although our tool places no restriction on the number of sequences to be used in the analysis, the user's web browser and hardware could be a limiting factor to visualize large SVG output files. TreeDomViewer was used to visualize a set of 530 Receptor-Like Proteins (RLP) obtained from the arabidopsis genome-wide survey of RLPs without any problem on a standard PC or Mac (data not shown).

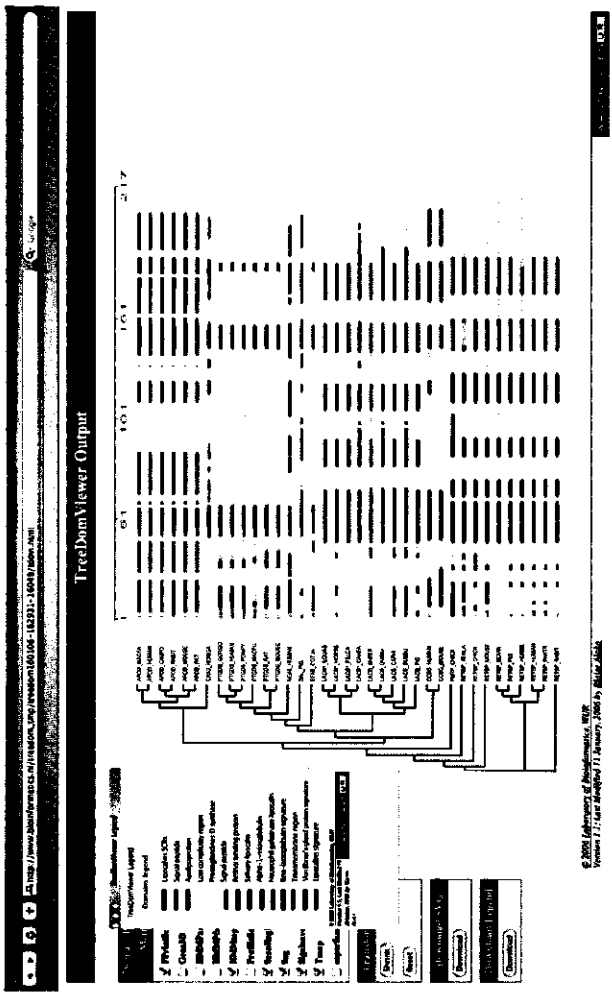


Figure 3: This figure illustrates the default SVG output of 37 lipocalin family members from different species. Shown in red are the main Structurally Conserved Residues (SCRs) that characterize the lipocalins. Insert shows TreeDomViewer domain legend (which appears as a separate pop-up).

Future plans

We intend to broaden the scope of TreeDomViewer by incorporating secondary structure prediction in the visualization as well as presenting (offering) TreeDomViewer as a BioMOBY (Wilkinson et al. (17)) web-service to the scientific community. Furthermore we plan to improve TreeDomViewer performance by expanding the distributed network of cluster mirrors.

Conclusion

TreeDomViewer is a biological web-based tool combining in one picture protein information on phylogenetic and structural information. As such it provides information about the relatedness of proteins and protein families, and thus adds support for inferring function of gene products, in particular when sequence identity is low. TreeDomViewer therefore helps in any phylogenetic analysis resolving both the relationship among different group members and the relationship between groups, based solely on the aligned domain structure of each participant.

Acknowledgements

The authors wish to thank Pieter Neerincx for testing the tool on Mac OS X and providing valuable tips for preparing the figures. This project was (co) financed by the Centre for BioSystems Genomics (CBSG) which is part of the Netherlands Genomics Initiative / Netherlands Organization for Scientific Research.

References

1. Constantinesco, F., Forterre, P., Koonin, E.V., Aravind, L. and Elie, C. (2004) A bipolar DNA helicase gene, herA, clusters with rad50, mre11 and nurA genes in thermophilic archaea. *Nucl. Acids Res.*, **32**, 1439-1447.
2. Bujnicki, J.M. (2004) In Gross, H. J. (ed.), *Nucleic Acids and Molecular Biology*. 1 ed. Springer, Vol. 15, pp. 146-148.
3. Jones, D.T., Taylort, W.R. and Thornton, J.M. (1992) A new approach to protein fold recognition. **358**, 86-89.
4. Pevsner, j. (2003) *Bioinformatics and functional genomics*. 1 ed. Wiley-Liss.
5. Todd, A.E., Orengo, C.A. and Thornton, J.M. (2002) Sequence and Structural Differences between Enzyme and Nonenzyme Homologs. *Structure*, **10**, 1435-1451.
6. Von Heijne, G. (1986) A new method for predicting signal sequence cleavage sites. *Nucleic Acids Res*, **14**, 4683-4690.
7. Milpetz, F., Argos, P. and Persson, B. (1995) TMAP: a new email and WWW service for membrane-protein structural predictions. *Trends in Biochemical Sciences*, **20**, 204-205.
8. Tusnady, G.E. and Simon, I. (2001) The HMMTOP transmembrane topology prediction server. *Bioinformatics*, **17**, 849-850.
9. Zdobnov, E.M. and Apweiler, R. (2001) InterProScan - an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, **17**, 847-848.
10. Eisen, J.A. and Wu, M. (2002) Phylogenetic Analysis and Gene Functional Predictions: Phylogenomics in Action. *Theoretical Population Biology*, **61**, 481-487.
11. Rice, P., Longden, I. and Bleasby, A. (2000) EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics*, **16**, 276-277.
12. Thompson, J., Higgins, D. and Gibson, T. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucl. Acids Res.*, **22**, 4673-4680.
13. Saitou, N. and Nei, M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*, **4**, 406-425.

14. Felsenstein, J. (2002) PHYLIP (Phylogeny Inference Package) Version3.6a3 Distributed by the author. Department of Genome Sciences, university of Washington, Seattle.
15. Ganfornina, M.D., Gutierrez, G., Bastiani, M. and S, D. (2000) A Phylogenetic Analysis of the Lipocalin Protein Family. *Mol Biol Evol*, **17**, 114-126.
16. Flower, D.R., North, ACT. Attwood, T. K. (1993) Structure and sequence relationships in the lipocalins and related proteins. *Protein Sci*, **2**, 753-761.
17. Wilkinson MD, L.M. (2002) BioMOBY: An open source biological web services proposal. *Briefings in Bioinformatics*, **3**, 331-341.

Chapter 3

CoPub Mapper: mining MEDLINE based on search term co-publication

Blaise T. F. Alako[1], Antoine Veldhoven[2], Sjozef van Baal[3], Rob Jelier[4], Stefan Verhoeven[1], Ton Rullmann[1], Jan Polman[1], and Guido Jenster[2*]

[1]Department of Molecular Design & Informatics, Organon NV, P.O. Box 20, 5340 BH Oss, The Netherlands.
[2]Department of Urology, Erasmus MC, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands.
[3]Department of Genetics, Erasmus MC, Rotterdam, The Netherlands.
[4]Department of Medical Informatics, Erasmus MC, Rotterdam, The Netherlands.

BMC Bioinformatics (2005), 6, 51
(With permission)

Abstract

Background: High throughput microarray analyses result in many differentially expressed genes that are potentially responsible for the biological process of interest. In order to identify biological similarities between genes, publications from MEDLINE were identified in which pairs of gene names and combinations of gene name with specific keywords were co-mentioned.

Results: MEDLINE search strings for 15,621 known genes and 3,731 keywords were generated and validated. PubMed IDs were retrieved from MEDLINE and relative probability of co-occurrences of all gene-gene and gene-keyword pairs determined. To assess gene clustering according to literature co-publication, 150 genes consisting of 8 sets with known connections (same pathway, same protein complex, or same cellular localization, etc.) were run through the program. Receiver operator characteristics (ROC) analyses showed that most gene sets were clustered much better than expected by random chance. To test grouping of genes from real microarray data, 221 differentially expressed genes from a microarray experiment were analyzed with CoPub Mapper, which resulted in several relevant clusters of genes with biological process and disease keywords. In addition, all genes versus keywords were hierarchical clustered to reveal a complete grouping of published genes based on co-occurrence.

Conclusions: The CoPub Mapper program allows for quick and versatile querying of co-published genes and keywords and can be successfully used to cluster predefined groups of genes and microarray data.

Background

High throughput microarray analysis has made it possible to analyze the mRNA expression of most if not all human genes simultaneously [1;2]. The data generated from these analyses are overwhelming since hundreds of interesting differentially expressed genes can be identified in a single assay. Knowledge on expression levels of genes in different systems is useful, but does not directly answer biologically relevant questions, such as: What is the gene function? Where is the gene located within the genome? Where is the protein located within the cell? Most important is the answer to the question whether genes identified in microarray experiments have something in common, such as, are multiple genes part of a single biological pathway or proteins part of a protein complex? The public database, which contains much of the relevant information to answer these questions, is MEDLINE. Therefore, mining the MEDLINE database for all information on a set of genes of interest to extract and evaluate their co-occurrences with biological keywords and other genes, could reveal biologically relevant pathways [3-6].

The most widely used methodology to identify genes and proteins in text is by thesaurus-based concept extraction. Using a predefined gene name list, text phrases are compared to the thesaurus for matching. Complications for gene name thesauri are variations in full name spelling, use of abbreviations (gene symbols), the large number of synonyms (different name but same gene) and homonyms (same name but meaning different genes or unrelated concepts) [7;8]. Particularly homonyms in the form of abbreviations and acronyms create a serious problem of false positive assignment of a gene to a particular concept [9-13]. A complementary approach for gene/protein identification is "named entity recognition" in which a program learns to recognize concepts from text [14-16]. Due to the enormous synonym and homonym problems, named entity recognition encounters difficulties in achieving high performance gene name identification. A next step in text mining is linking of different concepts (such as gene names and keywords) that are identified. In the simplest method, co-occurrence of two concepts within the document can be used as an indication of linkage. Extensions of co-occurrence can include (i) the number of times a concept is found, (ii) how close concepts are to one another, such as, within a single sentence, and (iii) not just two, but the weighed combination of all concepts within a document. More sophisticated fact extraction methods can also retrieve information on the type of relationship between two concepts. Natural language processing (NLP) grammatically parses whole sentences to identify verbs and other connecting phrases that describe the correlation between concepts [3;4;6;17]. A third step in text mining takes linked concepts and groups them according to their co-occurrence and relationships. Again, this can be performed by simple clustering of the co-occurrence of pairs of concepts as well as complex multi-dimensional classification using weighed concept combinations [18;19]. This type of clustering of, for example, differentially expressed genes from a microarray experiment, can disclose, summarize, and visualize published knowledge, but can also be utilized for novel information discovery [5;20]. Although progress is being made in higher order literature processing, text mining applications in the field of genomics are mainly thesaurus and co-occurrence based. Such programs and methods to identify potential functional correlations between genes have been described [21-33]. Each of these applications has its unique advantages and limitations, showing the broad range of needs for text mining as well as the numerous extraction, linking, and discovery methods feasible.

We set out to create a well annotated and curated open source gene list including full names, symbols and aliases and a regular expression-based search method to identify genes in text databases such as MEDLINE. In addition to the gene thesaurus, specific keyword lists were generated for co-occurrence analyses. For each concept, PubMed identifiers (IDs) from MEDLINE documents containing the concept were extracted, all gene-gene and gene-keyword co-occurrence pairs identified and stored in a database for fast co-occurrence retrieval. This database can be mined using single or batches of concepts to retrieve co-occurrences that form the input in clustering programs to group genes and keywords according to their similarity in co-publications. The program, database and all thesauri are freely available and can be adapted to include updates, new thesauri, and search methods.

Implementation

Human gene thesaurus

A human gene thesaurus was compiled from the Affymetrix HG_U95 / HG_U133 and HUGO gene annotations (HG_U95 / HG_U133 annotation files from 2002) [34] [8] (Table 1).

Thesaurus	Data Source	Number of terms	Number of terms with MEDLINE hits	Total number of MEDLINE citations
Gene	Affymetrix HG_U95-133 HUGO	15,621	10,700	5,932,448
Molecular Function	Gene Ontology	962	851	6,616,546
Cellular Component	Gene Ontology	218	196	1,890,561
Biological Process	Gene Ontology	767	621	3,455,950
Diseases	Karolinska Institute	1475	1444	6,099,280
Tissues	National Library of Medicine	309	307	9,083,831

Table 1: CoPub Mapper gene and keyword database information. Gene names, symbols and aliases were retrieved from Affymetrix HG_U95 / HG_U133 [54] and the HUGO databases [55]. The keyword thesauri include the three Gene Ontology subsections [41], diseases [56] and tissues/organs [57].

In total, 15,621 annotated genes were included of which most gene descriptions consist of one or more full names, the gene symbol, and their aliases. The typical HUGO and Affymetrix full gene name descriptions contain commas, semicolons and often alternative names in parenthesis, which makes this description an inadequate direct search term. Full names were processed by replacing the commas and semicolons with the Boolean "AND" operator (Figure 1).

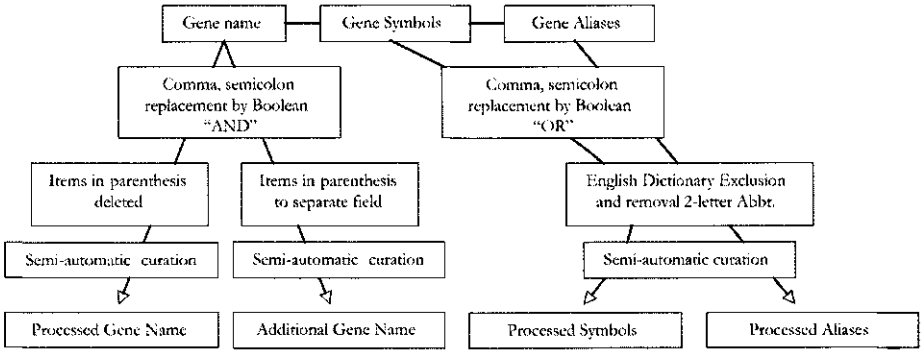


Figure 1: Flow diagram of the processing and curation of the gene names, symbols and aliases. Gene names, symbols and aliases were retrieved from Affymetrix HG_U95 / HG_U133 and the HUGO databases.

All terms included in parentheses were deleted from "gene-level name" and placed in a separate field named "gene-level additional description". Both fields were semi-automatically curated to remove common words (such as protein, family, hypothetical, functional, human, tissue, yeast, etc), misspellings, and insert Boolean "OR" in case synonyms are described. From gene symbols and aliases fields, commas and semicolons separators were replaced by the Boolean "OR" operator. Two-letter symbols and aliases were removed from the thesaurus and all other abbreviations were compared to an English dictionary [35] to remove common English words (such as "AND", "CELL", etc.). The Microsoft Excel spreadsheet program was used for generating and curating gene thesaurus files and, as described by Zeeberg et al [36], conversion problems were encountered and when identified, manually corrected.
Semi-automatic stemming was performed on "gene-level name" and "gene-level additional description" fields by removing numbers, letters, and phrases like "alpha", "member", "type", "class", etc. This resulted in a stem-level gene name description. Although the current version of CoPub Mapper does not take this stem-level into account, these fields are part of the gene thesaurus and freely available.

Keyword thesauri
In total, five different keyword thesauri were compiled including the Gene Ontology "biological process", "cellular component", and "molecular function", as well as "diseases" and "tissues" (Table 1). In the disease thesaurus, commas were replaced with the Boolean "OR" operator. All keyword databases were manually curated to remove terms too specific or too common.

MEDLINE concept extraction and curation
The full MEDLINE baseline XML files (until January 2004) were obtained from the National Library of Medicine [37], extracted to small text files containing title, abstract

and substances using BioPerl API. The title, substance and abstract fields from MEDLINE records from 1966 to January 2004 were searched for the presence of different case-insensitive gene and keyword concepts using Perl compatible regular expressions (PCRE). For the gene-level name descriptions the characters "][.-)(,:;" and space were allowed preceding and following the gene-level name description and also an optional "s" was permitted to follow the name. Any space in the gene-level name description was allowed to be a space or a dash. The same regular expressions were applied to the gene name stem-level descriptions, except that, the description could also be followed by any single letter or a number between 0 and 99. Gene symbols and aliases could be preceded and followed by the characters "][.-)(,:;" and space. After the first two characters, the presence of a dash was allowed in between the characters of the symbols and aliases (to take, for example, both "bcl2" and "bcl-2" into account). The concepts of the keyword files could be preceded and followed by the characters "][.-)(,:;" and space. In addition, "s" and "'s" were allowed to follow the disease concept. As for the gene-level name descriptions, a dash was allowed to be present between the words of a keyword concept. Per annotated gene or keyword, the PubMed IDs of MEDLINE records in which the concept was identified were stored in a MySQL database.

In order to identify potential problem concepts, 50 genes and 50 keywords with the highest number of PubMed IDs were manually inspected and curated if appropriate. In addition, a random selection of genes and all keywords that gave less than 2 MEDLINE hits were examined and this evaluation was used to optimise the thesauri and regular expressions search strategy described above.

To address the homonym issue, a correction was made for possible discrepancies between a parenthesised gene symbol and its expected name. All abbreviations in parenthesis in MEDLINE abstracts were retrieved in combination with 4 preceding words. In total, 1,105,669 MEDLINE records were identified where the abbreviation matched a gene symbol or alias. For all these records, 4 words preceding the abbreviation were compared to the gene-level name description of that particular gene. If none of the words resembled partly the gene name, the PubMed ID was removed from that particular gene's PubMed ID list. Using this method, 603,580 records were deleted from the gene hit database resolving part of the gene-unrelated concept homonym problems. Manual inspection of 173 random records revealed that, extrapolated, 79 % of the 603,580 records was correctly removed, while 7 % of the 502,089 non-removed records should have been deleted.

In our examination of genes with the highest number of PubMed IDs and our first CoPub Mapper analyses, we noticed a distinct contamination of records identifying gene symbols and aliases by abbreviation used for cell lines (such as PC3 which is an alias for 3 genes as well as a prostate cancer cell line). Since full names of cell line abbreviations are rarely put in writing, the homonym correction did not eliminate these discrepancies. A list of cell line names was retrieved [38] and gene symbols and aliases that fitted a cell line name were further processed. From 106 genes that included one of the cell line homonym names, all MEDLINE records were deleted in which the cell line name was mentioned without the presence of the stem-level gene name. In total, 100,213 PubMed IDs were eliminated. A manual inspection of 78 randomly chosen records showed that 87 % were correctly removed.

Database set-up and CoPub Mapper program

A file was generated that contains a unique query ID and the probeset IDs, UniGene (combination of Aug 2002 and Oct 2003 builds) and RefSeq identifiers for each of the individual 15,621 entries in the gene thesaurus (alias_affygene). In addition, a file with the gene name, symbol and aliases and unique query ID was created (query_affygene).

The retrieved PubMed IDs from each field (gene names, symbols and aliases) of the 15,621 unique gene thesaurus query IDs were non-redundantly combined into a MySQL database (lit_affygene) and a separate data-file (litstat_affygene) in which the number of PubMed IDs per query was counted. Furthermore, the PubMed IDs from the keyword thesauri were per concept stored (query_*keyword*, lit-*keyword* and litstat_*keyword*). Per gene-gene pair and gene-keyword pair, overlaps in PubMed IDs were identified and separately stored in the database (pair_*keyword*_affygene). From these paired files, a pairstat file was generated containing the number of PubMed IDs of each concept, the number of overlapping PubMed IDs between the two concepts and a relative score. The relative score is based on the mutual information measure and was calculated as

$$S = P_{AB}/P_A * P_B$$

in which P_A is the number of hits for concept A divided by the total number of PubMed IDs, P_B is the number of hits for concept B divided by the total number of PubMed IDs, and P_{AB} is the number of co-occurrences between concepts A and B divided by the total number of PubMed IDs. The relative score is produced as a log10 conversion and in the batch search option in a 1-100 scaled log10 conversion:

$$R = {}^{10}\log S$$

and the scaled log transformed relative score:

$$R' = 1 + 99 * (R - Rmin) / (Rmax - Rmin)$$

where Rmin and Rmax are the lowest and highest R values in each pairstat file, respectively.

The CoPub program was generated in Python and runs as a web-based application (CGI script). The text output of a batch search can be saved and imported into a clustering program such as Cluster [39] and SpotFire (Spotfire, Göteborg, Sweden). The HTML output of "number of hits", "relative score", and batch search results are hyperlinked to the MEDLINE database at the European Bioinformatics Institute [40] for direct manuscript retrieval.

Performance evaluation using ROC (receiver operating characteristics) curves

In order to investigate whether the CoPub Mapper output could group genes according to their MEDLINE co-occurrence profile, 8 different groups of genes were defined based on common gene ontology (GO) terms [41], the BRCA1 BioCarta pathway [42], or a microarray experiment (Table 2).

In the UniGEM V microarray experiment, the gene expression profile of prostate stroma cells was compared to prostate epithelial cells [43]. A set of 28 annotated genes, higher expressed in epithelial cells as compared to stromal cells (more than 2-fold) was randomly selected.

The 150 genes from the eight selected gene groups are pooled into one set. The selected genes were entered into CoPub Mapper to generate the co-occurrence matrix of relative scores of genes versus genes and genes versus the 5 different keyword thesauri. Relative scores were only generated in case more than 2 co-publications occurred per concept-concept pair. The genes versus genes matrix was hierarchical clustered and visualised using Cluster and TreeView [39] (Figure 2).

Test groups	#Genes	Source
smooth muscle contraction	12	GO (Biological Process)
acetyltransferase	18	GO (Molecular Function)
nuclear pore	15	GO (Cellular Component)
nucleosome	17	GO (Cellular Component)
ubiquitin	24	GO (Molecular Function)
hypoxia	26	GO (Biological Process)
BRCA1	11	BioCarta
Epithelial-specific genes	27	UniGEM V microarray: stroma vs epithelial cells

Table 2: CoPub Mapper test groups. Eight groups of genes with a common function, process, cellular location, or microarray expression profile, were defined from gene ontology (GO), BioCarta, or a microarray experiment. The genes used for CoPub Mapper analysis were randomly selected from larger sets of genes part of the 8 different groups.

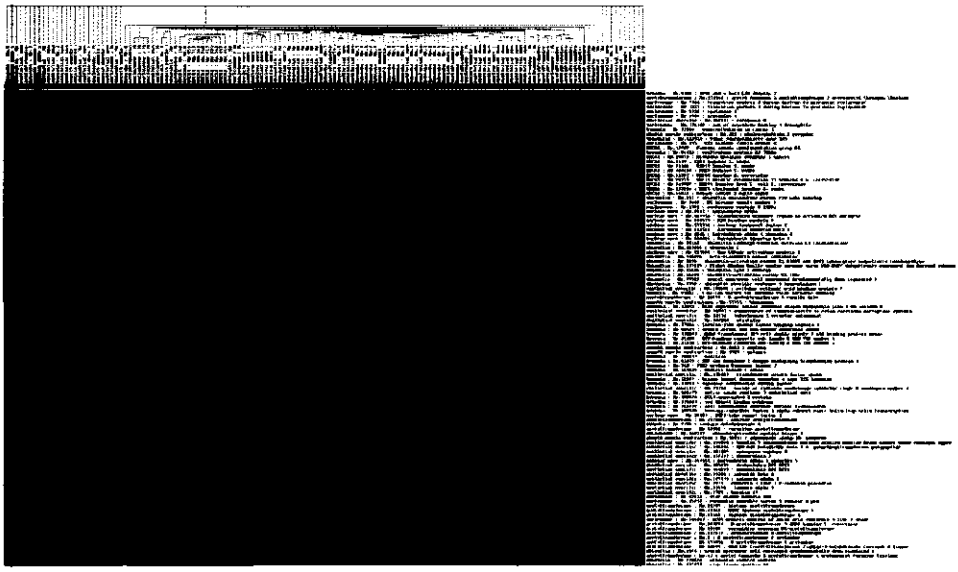


Figure 2: Clustered view of gene co-occurrences among a collection of 8 groups of selected genes. Of the 150 genes, the relative scores of co-occurrences were calculated and clustered using hierarchical clustering. A co-occurrence was only taken into account when at least two articles mention the gene-gene pair. Using this criterion, 45 genes did not co-publish with any of the other 149 genes. To which group (Table 2) a gene belongs to is indicated in the right part of the figure. Image contrast in TreeView was set at 50. Scaled (1-100) relative scores are represented in a red spectrum with bright red being the highest score. A relative score of zero or no score are in black.

For a systematic evaluation of performance we applied Receiver Operating Characteristics (ROC) graphs and the area under the ROC curve (AUC) as an outcome measure. To use this method all genes from the 8 subgroups are pooled into one set. To calculate an AUC for every gene we used the following procedure. A gene from the pooled set is selected as a seed. The seed is paired with all other genes in the set and non-centered Pearson correlation coefficients are calculated based on their co-occurrence profiles. The co-occurrence profile is one row of the co-occurrence matrix under investigation. The genes are ordered by their correlation coefficients, with the highest value at the first rank. To generate a ROC curve, the obtained ranking of the genes is viewed as the outcome of a classifier. For a seed, genes from the same subgroup are called positives and all other genes are called negatives. ROC curves are two-dimensional graphs in which the true-positive (TP) rate is plotted against the false-positive (FP) rate. The TP rate is defined as correctly classified positives divided by all positives. The FP rate is defined as incorrectly classified negatives divided by all negatives. While running down the list, for every rank the true and false positive rate are calculated, by taking all encountered genes to be classified as positive and all not yet encountered genes as negative. The AUC of the ROC curve is calculated. The procedure is repeated until an AUC has been calculated for every gene in the pooled set. An average AUC is calculated per subgroup. The AUC measure varies between 0 and 1. Random ordering gives an AUC of 0.5 and an AUC of 1 represents perfect ordering, i.e. all positives are at the top of the list with no negatives in between, indicating perfect co-occurrence clustering of the genes in the subgroup [44].

Results

Validation of CoPub Mapper co-occurrence profiling

To validate the usefulness of the CoPub Mapper output, we evaluated how well genes with known relations could be grouped according to their MEDLINE co-occurrence profile. As shown in Figure 2, partial clustering of the initial 8 groups occurred upon their gene-gene co-occurrence profile evaluation. To quantify this grouping, ROC (receiver operating characteristics) curves were generated and the AUCs (Area Under Curve) for each gene calculated. In Figure 3, the median AUCs $\pm$ SD of the genes per group are depicted. Most of the 8 groups and in particular the BRCA1-associated genes clustered well together in the gene-keyword comparisons (median AUC of 0.93 ± 0.07). The ubiquitin-associated genes performed worst (median AUC of 0.6 ± 0.11). With respect to the thesaurus selection, the overall clustering of the 8 groups using the "genes versus genes self" comparison, performed best with an average AUC of 0.76 ± 0.13. The "genes versus diseases" and "genes versus tissues" comparisons were for many of the 8 groups not resulting in clustering higher than expected by random chance. In other words, from co-publication analysis of genes with disease or tissue keywords, the commonality between the genes, as defined by the 8 groups, could rarely be traced (Figure 3).

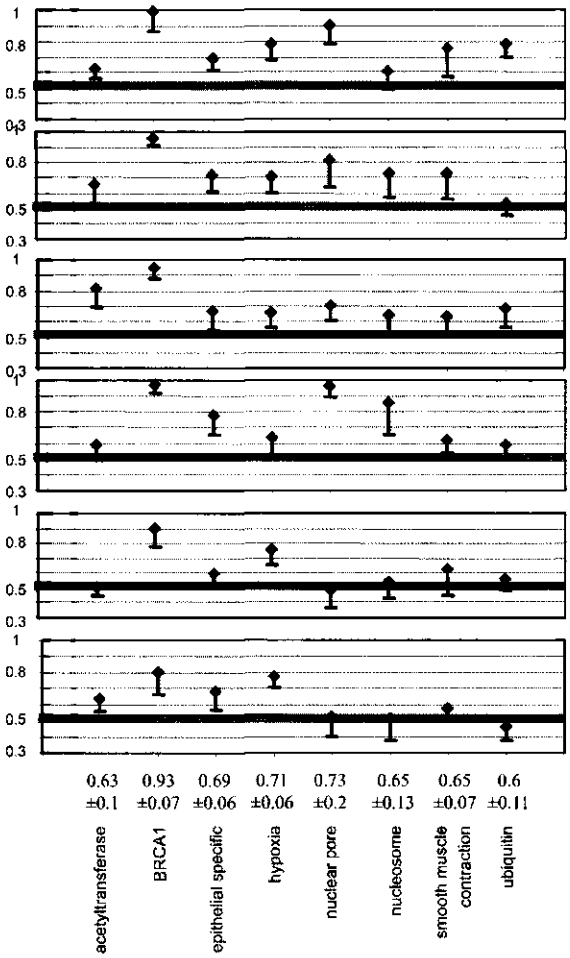


Figure 3: Receiver operating characteristics (ROC) of the 8 selected groups of genes to quantify their coherence upon clustering of literature co-occurrences. Co-occurrences of the 150 genes were determined with the genes themselves, or the 5 different keyword thesauri. A co-occurrence was only taken into account when at least two articles mention the gene-gene or gene-keyword pair. The co-occurrence matrixes were Pearson correlation clustered and the distances between genes determined. For each gene, it was determined whether the next closest clustered gene was a group member. Genes from the same group were scored as true positive and any other gene as false positive to generate a ROC curve. For each gene, the area under the ROC curve (AUC) was determined and the median of all the group members per group ± SD depicted. Scaling is from an AUC of 0.3 to 1. An AUC of 0.5, representing a random ordering is highlighted with a thick line.

As shown in Table 2, six groups of genes were selected based on gene ontology keywords, using two from each of the annotation trees (biological process, molecular function, and cellular component). As expected and without exception, the AUC of the 6 groups of genes was higher using their corresponding GO-derived thesaurus compared to using the other two GO-derived thesauri. For example, the molecular function annotated group of "acetyltransferases" was clustered best using the "genes versus molecular function" co-publication comparison (AUC of 0.81 as compared to 0.65 using the biological process thesaurus and 0.59 using the cellular component thesaurus). This shows that the selection of keywords for co-occurrence analysis is an important determinant in optimal text-based grouping of genes.

Microarray analysis using CoPub Mapper

In order to validate the CoPub Mapper program with real microarray data, a set of differentially expressed genes was selected from a comparison between ovaries of healthy women and women suffering from Poly Cystic Ovary Syndrome (PCOS) [45]. PCOS is characterized by a combination of chronic anovulation, hyperandrogenism and cysts in ovaries and is the most common cause of anovulatory infertility. Also hyperinsulinemia and obesity can be observed in many PCOS patients [46;47].

A set of 230 dysregulated DNA fragments representing 189 genes were used as input for CoPub Mapper (see Table 1 in [45]). Gene-keyword pairs were obtained from biological processes and diseases. Relative scores were only generated in case 3 or more co-publications occurred per gene-keyword pair. From these 189 genes, 104 were annotated and had at least 3 co-publications with one of the keywords. Resulting matrices were exported as text files and opened and merged in Spotfire. Hierarchical clustering was used to group genes and keywords. Figure 4 shows that subsets of genes form clusters with subsets of biological processes and diseases. Zooming in on these clusters confirms the relation of certain genes with e.g. PCOS, diabetes, obesity, gametogenesis, immune response. Characterization of all clusters revealed known and unknown relations of these PCOS dysregulated genes with biological processes and diseases.

Single Gene-Keyword extraction

The CoPub Mapper includes an option to query the database for all genes and keywords co-published with a single gene of interest. In addition, a keyword of interest can be selected and all genes with 2 or more co-occurrences can be extracted. As examples, the top ten genes (Table 3) and top ten diseases (Table 4) co-published with the androgen receptor are shown.

An assessment of the 2 lists identified the puromycin-sensitive aminopeptidase gene (NPEPPS) as an example of a homonym (Table 3, fourth gene). The PSA alias of NPEPPS is mainly used to specify prostate specific antigen. The prostate specific antigen gene (KLK3) is regulated by the androgen receptor and correctly found many times to be co-published with the androgen receptor (Table 3, second gene). Due to the homonym curation described in the Systems and Methods section, the number of co-occurrences of the androgen receptor with NPEPPS (246) is lower than with KLK3 (414). Before homonym curation, NPEPPS and KLK3 had 634 and 635 co-publications with the androgen receptor, respectively. The top ten list of diseases co-published with the androgen receptor (Table 4) is a near perfect reflection of the known diseases associated with androgen receptor activity and aberrations.

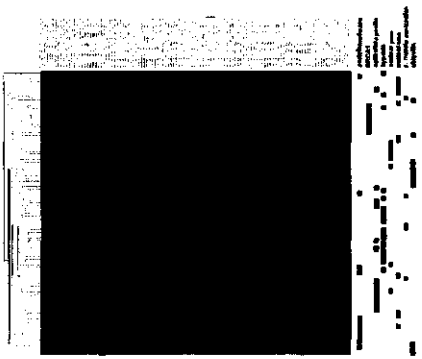


Figure 2: Clustered view of gene co-occurrences among a collection of 8 groups of selected genes. Of the 150 genes, the relative scores of co-occurrences were calculated and clustered using hierarchical clustering. A co-occurrence was only taken into account when at least two articles mention the gene-gene pair. Using this criterion, 45 genes did not co-publish with any of the other 149 genes. To which group (Table 2) a gene belongs to is indicated in the right part of the figure. Image contrast in TreeView was set at 50. Scaled (1-100) relative scores are represented in a red spectrum with bright red being the highest score. A relative score of zero or no score are in black.

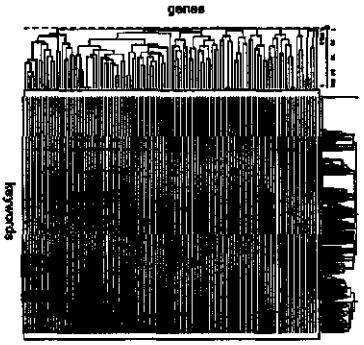


Figure 4: Hierarchical clustering of literature co-occurrences of 104 genes (rows) versus 761 biological processes and diseases (columns). A co-occurrence was only taken into account when at least three articles mention the gene-keyword pair. Hierarchical clustering of CoPub Mapper results using genes differentially expressed in PCOS ovaries. From 221 regulated genes 104 genes contain a gene name, symbol or alias and produce a gene-keyword pair with biological processes or diseases. 104 modulated genes resumed 761 keywords denoting biological processes or diseases. Hierarchical clustering was performed using Spotfire using the Complete Linkage method and Correlation as Similarity Measure. Several subclusters were identified shown here with blue boxes; between parentheses the number of genes in a cluster. A: PCOS, Obesity, Insulin Resistance (4); B & D: Gametogenesis (5&8); C: Cell adhesion, Angiogenesis (19); E & H: Immune response, Inflammation (14&11); F: Cancer, Cell growth, Differentiation (32); G: Inflammatory diseases (6).

Gene Name	Gene Symbols	Gene Alias	Pmid Hits
progesterone receptor	PGR	NR3C3	605
kallikrein 3, prostate specific antigen, prostate specific antigen	KLK3	PSA	414
(nuclear receptor subfamily 3, group C, member 1), (glucocorticoid receptor)	NR3C1	GCR, GRL	389
aminopeptidase puromycin sensitive	NPEPPS	MP100, PSA	246
sex hormone-binding globulin	SHBG	ABP	179
gonadotropin-releasing hormone 1, leutinizing-releasing hormone	GNRH1	GNRH, GRH, LHRH, LNRH	157
prolactin	PRL		131
insulin	INS		125
epidermal growth factor, beta-urogastrone	EGF	URG	123
tumor protein p53	TP53	P53	94

Gene Pairs

Total Number of Gene Pairs found: 1436

	Gene Name1	Gene Symbols1		Gene Name2			
1	androgen receptor, dihydrotestosterone receptor		DHTR, NR3C4	progesterone receptor	PGR	NR3C3	605
2	androgen receptor, dihydrotestosterone receptor		DHTR, NR3C4	kallikrein 3, prostate specific antigen, prostate specific antigen	KLK3	PSA	414
3	androgen receptor, dihydrotestosterone receptor		DHTR, NR3C4	(nuclear receptor subfamily 3, group C, member 1), (glucocorticoid receptor)	NR3C1	GCR, GRL	389
4	androgen receptor, dihydrotestosterone receptor		DHTR, NR3C4	aminopeptidase puromycin sensitive	NPEPPS	MP100, PSA	246
5	androgen receptor, dihydrotestosterone receptor		DHTR, NR3C4	sex hormone-binding globulin	SHBG	ABP	179
6	androgen receptor, dihydrotestosterone receptor		DHTR, NR3C4	gonadotropin-releasing hormone 1, leutinizing-releasing hormone	GNRH1	GNRH, GRH, LHRH, LNRH	157
7	androgen receptor, dihydrotestosterone receptor		DHTR, NR3C4	prolactin	PRL		131
8	androgen receptor, dihydrotestosterone receptor		DHTR, NR3C4	insulin	INS		125
9	androgen receptor, dihydrotestosterone receptor		DHTR, NR3C4	epidermal growth factor, beta-urogastrone	EGF	URG	123
10	androgen receptor,		DHTR,	tumor protein p53	TP53	P53	94

Figure 4: Hierarchical clustering of literature co-occurrences of 104 genes (rows) versus 761 biological processes and diseases (columns). A co-occurrence was only taken into account when at least three articles mention the gene-keyword pair. Hierarchical clustering of CoPub Mapper results using genes differentially expressed in PCOS ovaries. From 221 regulated genes 104 genes contain a gene name, symbol or alias and produce a gene-keyword pair with biological processes or diseases. 104 modulated genes returned 761 keywords denoting biological processes or diseases. Hierarchical clustering was performed using Spotfire using the Complete Linkage method and Correlation as Similarity Measure. Several subclusters were identified shown here with blue boxes; between parentheses the number of genes in a cluster. A: PCOS, Obesity, Insulin Resistance (4); B & D: Gametogenesis (5&8); C: Cell adhesion, Angiogenesis (19); E & H: Immune response, Inflammation (14&11); F: Cancer, Cell growth, Differentiation (32); G: Inflammatory diseases (6).

Gene Name	Gene Symbols	Gene Alias	Pmid Hits
progesterone receptor	PGR	NR3C3	605
kallikrein 3, prostate specific antigen	KLK3	PSA	414
nuclear receptor subfamily 3, group C, member 1; glucocorticoid receptor	NR3C1	GCR, GRL	389
aminopeptidase puromycin sensitive	NPEPPS	MP100, PSA	246
sex hormone-binding globulin	SHBG	ABP	179
gonadotropin-releasing hormone 1, leutinizing-releasing hormone	GNRH1	GNRH, GRH, LHRH, LNRH	157
prolactin	PRL		131
insulin	INS		125
epidermal growth factor, beta-urogastrone	EGF	URG	123
tumor protein p53	TP53	P53	94

Table 3: CoPub Mapper single gene pair output. Output of the "Single Gene Pair Mapper" in which the top ten genes co-published with the androgen receptor are listed according to number of co-publications (Pmid hits).

Keywords	Number of hits	log10 Relative Score
Androgen-Insensitivity Syndrome	229	3.07
Kennedy Disease	21	2.56
Muscular Atrophy Spinal	133	2.12
Prostate Cancer	932	1.93
Gynecomastia	59	1.88
Hypospadia	81	1.79
Sex Chromosome Aberrations	2	1.78
Hirsutism	76	1.78
Robinow Syndrome	2	1.71
X-Linked Myotubular Myopathy	2	1.65

Table 4: CoPub Mapper single gene biological concept output. Output of the "Single Gene Biological Term Mapper" in which the top ten diseases co-published with the androgen receptor are listed according to their relevance score.

In Table 5, the top ten genes are listed that are most often co-published with the keyword "prostate cancer". Again, the incorrect identification of NPEPPS in 4507 MEDLINE entries is due to the PSA homonym.

Gene name	Gene Symbols	Gene Aliases	Number of hits	log10 Relative Score
kallikrein 3, prostate specific antigen	KLK3	PSA	6628	2.55
aminopeptidase puromycin sensitive	NPEPPS	MP100,PSA	4507	2.57
androgen receptor, dihydrotestosterone receptor		DHTR,NR3C4	932	1.93
acid phosphatase,prostate	ACPP		546	2.22
gonadotropin-releasing hormone 1,leutinizing-releasing hormone	GNRH1	GNRH,GRH,LHRH,LNRH	522	1.24
tumor protein p53	TP53	P53	431	0.96
B-cell CLL/lymphoma 2	BCL2		346	1.17
insulin	INS		318	0.05
epidermal growth factor,beta-urogastrone	EGF	URG	251	0.72
cyclin-dependent kinase inhibitor 1A	CDKN1A	CAP20, CDKN1, CIP1, MDA-6, P21, SDI1, WAF1	190	0.98

Table 5: CoPub Mapper single gene biological concept output. Output of the "Single Gene Biological Term Mapper" in which the top ten genes co-published with the prostate cancer disease-keyword are listed according to number of co-publications.

Meta-analysis: all genes versus keywords

In order to provide a summary of all gene-keyword co-occurrences, CoPub Mapping was performed using all 15,621 annotated genes as input in the different gene-keyword thesauri co-occurrence comparisons. Relative scores were only computed if in at least two articles a co-occurrence was observed. Elimination of single gene-keyword co-publications was carried out to eradicate non-reproduced findings and to make the large matrices manageable. A second selection was made to eliminate genes which included only low relative scores. Many genes have multiple co-publications with very common keywords such as "cancer" (disease thesaurus), "cytoplasm" (cellular component thesaurus), etc.. If not functionally relevant, these co-occurrences have typically a low relevance score. Genes with only low relevance scores were eliminated by removing those genes that did not have 1 or more scaled relevance scores of more than a threshold (between 39 and 52) in which 20 % of genes were eliminated. The hierarchical clustered genes-diseases co-publication matrix is displayed in Figure 5.

5626 genes (rows) versus 1275 diseases (columns) were grouped according to their co-publication profiles. The enlarged section shows the amount of detail present in the matrix (Figure 5B). The vertical lines in the matrix are caused by co-publication of almost all genes with very common disease keywords such as "cancer", "neoplasm", and "carcinoma". Horizontal lines are genes co-published with many diseases, such as "insulin", "interleukin 6", and "keratin 3A". If low relevance scores are masked by hiding values below 30 in TreeView or SpotFire, these streaks become less prominent.
Clustering and visualisation of only highly significant co-occurrences will result in discrete groups of genes and keywords as shown in Figure 6.

Stringent selection criteria were implemented including: (i) each gene had to be co-published with at least two different keywords with a relevance score of more than 50, and (ii) a co-occurrence must have been described in at least 3 publications per gene-keyword combination. From the 10,203 genes co-occurring with cellular component keywords, 1135 genes were retrieved using the stringent selection criteria mentioned above. As expected, these genes were clustered according to well-known cellular components of which some examples are depicted (Figure 6).

Keywords	Number of hits	log10(Relativ Score)
Androgen-Insensitivity Syndrome	229	3.07
Kennedy&&&Disease	21	2.56
Muscular Atrophy&&&Spinal	133	2.12
Prostate Cancer	932	1.93
Gynecomastia	59	1.88
Hypospadia	81	1.79
Sex Chromosome Aberrations	2	1.78
Hirsutism	76	1.78
Robinow Syndrome	2	1.71
X-Linked Myotubular Myopathy	2	1.65

Count	[redacted]	Gene Symbol(s)	Gene Aliase(s)	[redacted]	Number of hits	log10(Relative Score)
1	androgen receptor, dihydrotestosterone receptor		DHTR, NR3C4	Androgen-Insensitivity Syndrome	229	3.06501547504
2	androgen receptor, dihydrotestosterone receptor		DHTR, NR3C4	Kennedy&&Disease	21	2.55864138166
3	androgen receptor, dihydrotestosterone receptor		DHTR, NR3C4	Muscular Atrophy&&Spinal	133	2.11586003451
4	androgen receptor, dihydrotestosterone receptor		DHTR, NR3C4	Prostate Cancer	932	1.9266829564
5	androgen receptor, dihydrotestosterone receptor		DHTR, NR3C4	Gynecomastia	59	1.87903047224
6	androgen receptor, dihydrotestosterone receptor		DHTR, NR3C4	Hypospadia	81	1.78543865032
7	androgen receptor, dihydrotestosterone receptor		DHTR, NR3C4	Sex Chromosome Aberrations	2	1.77848408192
8	androgen receptor, dihydrotestosterone receptor		DHTR, NR3C4	Hirsutism	76	1.77737969272
9	androgen receptor, dihydrotestosterone receptor		DHTR, NR3C4	Robinow Syndrome	2	1.71354346163
10	androgen receptor, dihydrotestosterone receptor		DHTR, NR3C4	X-Linked Myotubular Myopathy	2	1.64659681199

Figure 5: Hierarchical clustering of literature co-occurrences of 5626 genes (rows) versus 1275 diseases (columns). A co-occurrence was only taken into account when at least two articles mention the gene-disease pair. Each gene had to have at least once a high (1-100 scaled) relevance score of >46. A: Overview of all 5626 genes and 1275 diseases. B: Enlargement of a small subsection of genes showing the amount of detail present in the CoPub Mapper analysis.

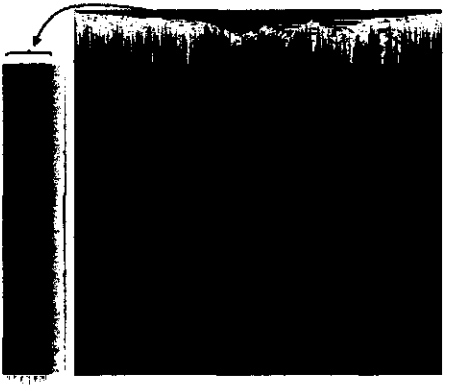
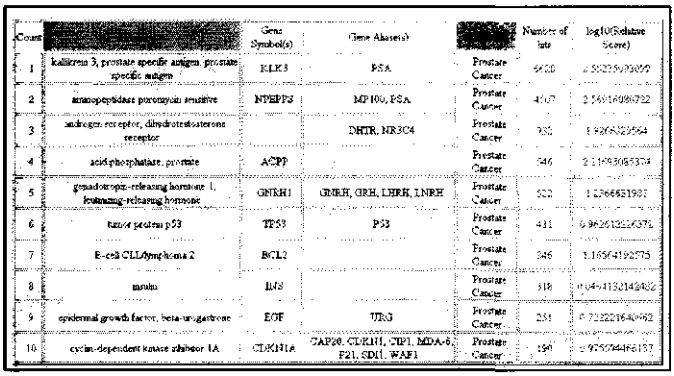


Figure 5: Hierarchical clustering of literature co-occurrences of 5626 genes (rows) versus 1275 diseases (columns). A co-occurrence was only taken into account when at least two articles mention the gene-disease pair. Each gene had to have at least once a high (1-100 scaled) relevance score of >46. A: Overview of all 5626 genes and 1275 diseases. B: Enlargement of a small subsection of genes showing the amount of detail present in the CoPub Mapper analysis.

Figure 6: Hierarchical clustering of literature co-occurrences of 1135 genes (rows) versus 177 cellular components (columns). A co-occurrence was only taken into account when at least three articles mention the gene-cellular component pair. Each gene had to have at least twice a high (1-100 scaled) relevance score of >50. Relative scores of less then 50 were masked in the TreeView program. Some of the cellular component concepts responsible for clustering of genes are indicated.

Gene name	Gene Symbols	Gene Aliases	Number of hits	log10(Relativ Score)
kallikrein 3,prostate specific antigen,prostate specific antigen	KLK3	PSA	6628	2.55
aminopeptidase puromycin sensitive	NPEPPS	MP100,PSA	4507	2.57
androgen receptor, dihydrotestosterone receptor		DHTR,NR3C4	932	1.93
acid phosphatase,prostate	ACPP		546	2.22
gonadotropin-releasing hormone 1,luteinizing-releasing hormone	GNRH1	GNRH,GRH,LHRH1,LNRH	522	1.24
tumor protein p53	TP53	P53	431	0.96
B-cell CLL/lymphoma 2	BCL2		346	1.17
insulin	INS		318	0.05
epidermal growth factor,beta-urogastrone	EGF	URG	251	0.72
cyclin-dependent kinase inhibitor 1A	CDKN1A	CAP20, CDKN1, CIP1, MDA-6, P21, SDI1, WAF1	190	0.98



Count		Gene Symbol(s)	Gene Aliases(s)		Number of hits	log10(Relative Score)
1	kallikrein 3, prostate specific antigen, prostate specific antigen	KLK3	PSA	Prostate Cancer	6628	2.5225393959
2	aminopeptidase puromycin sensitive	NPEPPS	MP100, PSA	Prostate Cancer	4507	2.5616980922
3	androgen receptor, dihydrotestosterone receptor		DHTR, NR3C4	Prostate Cancer	932	1.9266329564
4	acid phosphatase, prostate	ACPP		Prostate Cancer	546	2.2169305337
5	gonadotropin-releasing hormone 1, luteinizing-releasing hormone	GNRH1	GNRH, GRH, LHRH, LNRH	Prostate Cancer	522	1.2766621987
6	tumor protein p53	TP53	P53	Prostate Cancer	431	0.9626139216376
7	B-cell CLL/lymphoma 2	BCL2		Prostate Cancer	346	1.1656419257
8	insulin	INS		Prostate Cancer	318	0.0547132149482
9	epidermal growth factor, beta-urogastrone	EGF	URG	Prostate Cancer	251	0.7122221640962
10	cyclin-dependent kinase inhibitor 1A	CDKN1A	CAP20, CDKN1, CIP1, MDA-6, P21, SDI1, WAF1	Prostate Cancer	190	0.9750744468137

Figure 6: Hierarchical clustering of literature co-occurrences of 1135 genes (rows) versus 177 cellular components (columns). A co-occurrence was only taken into account when at least three articles mention the gene-cellular component pair. Each gene had to have at least twice a high (1-100 scaled) relevance score of >50. Relative scores of less then 50 were masked in the TreeView program. Some of the cellular component concepts responsible for clustering of genes are indicated.

Discussion

With the implementation of high-throughput technologies in many fields of research, problems have shifted from data gathering to data comprehension. Linking data from different sources, such as microarray expression data to biomedical text corpora, can assist in the disclosure, summary, and visualisation of knowledge. This is particularly valuable when from high throughput data, only a few items can be selected for further detailed low-throughput examination. Co-occurrence analysis of concepts using the MEDLINE literature database, is an effective tool to extract and categorize published knowledge. CoPub Mapper output was successfully used to cluster predefined groups of genes and resulted in a commonsensical clustering of PCOS microarray data. In addition, CoPub Mapper uncovered relationships between genes using single concept searches and provided an overall gene-keyword clustered summary of the literature. One obvious limitation of gene-driven text mining is the incomplete study and publication of all human genes. Out of approximately 30,000 human genes, we included 15,621 annotated genes of which 10,700 were mentioned at least once and 9,769 at least twice in MEDLINE. The use of human gene names, symbols and aliases does not necessarily mean a human-specific literature search. Many gene names and symbols are shared by other species as well.

The main advantages of CoPub Mapper above most other co-publication programs, are its modularity of keyword databases and the pre-calculated co-occurrences. Based on the results from the predefined groups of genes, the choice of keyword database made a substantial difference in clustering efficiency as determined by AUC calculations. Utilisation of a single joint thesaurus could counteract clustering due to inclusion of irrelevant non-discriminating keywords. Another illustration that keyword selection is an important issue is the prevalence of common keywords such as "cancer" (disease), "membrane" (cellular component), "metabolism" (biological process), "receptor" (molecular function), and "blood" (tissue). These keywords are co-published with nearly any gene of interest and were identified using CoPub Mapper. Although the relative score is generally low, these co-occurrences will influence the clustering process. Manual removal or stringent selection criteria before clustering can largely eliminate this potential bias. Addition of new keyword thesauri such as species, technologies, drugs, toxicology, pathology, etc. is feasible. Pre-calculation of co-publication of all possible gene-gene and gene-keyword pairs and storage in the pairstat data file, makes querying the database extremely efficient. Although the data are present, CoPub Mapper is not programmed for co-occurrence querying of more than 2 concepts. We are currently integrating CoPub Mapper into the Sequence Retrieval System (SRS) for multi-concept interrogation and direct linkage to other databases (such as microarray data, Gene Ontology, OMIM, SwissProt, LocusLink, UniGene, Ensembl, etc.) [48].

Comparing the gene expression profiles of normal versus PCOS ovaries has identified a large number of genes representing networks and pathways that are deregulated in PCOS. However, the gene names and symbols hardly ever point to specific signal transduction pathways. The relation of genes with their function, localization and context has been described in literature. Here we show that within the list of differentially expressed genes some are linked to PCOS, obesity, diabetes and gametogenesis. This is without surprise and easily explained [46;47]. Other genes are linked to cell proliferation, differentiation and cancer. Most of them were downregulated which correlates with the

observed arrest in growth and differentiation of follicles. Other clusters with no obvious link to PCOS may shed new light on the genes and pathways involved in the disease.

One of the major challenges associated with compiled heterogeneous text records such as MEDLINE, is correct gene recognition and assignment. The lack of consistent gene naming has resulted in a flood of synonyms and homonyms [7]. Although the synonym issue can be resolved by accumulating all different gene names and symbols, the correction for homonyms is still a daunting task. In order to include different spelling forms and the word context, we performed the text searches case insensitive and with predefined rules of regular expression.
The homonym problem consists of (i) different genes with identical gene name, symbol, or alias, and (ii), more frequently, a gene name, symbol or alias used for other terms than genes [9]. In the curated CoPub Mapper gene thesaurus, 1,286 of the 15,621 annotated genes (8.2 %) share a symbol or alias. In order to limit both aspects of the homonym problem, we (i) eliminated 2 letter symbols and aliases, (ii) deleted all symbols and aliases present in the English dictionary, (iii) manually curated terms with exceptionally high number of hits, (iv) corrected for cell line names, and (v) deleted records in which the preceding description of parenthesised symbols or aliases did not match the corresponding gene name. This last method has been used before to make an inventory of the homonym problem and provide strategies for correction, such as the one used here [9-13]. Although these measures effectively reduced the homonym problem, one will regularly encounter incorrect record assignment and invalid co-occurrence quotation using CoPub Mapper. Additional optimisation of the gene thesaurus might further reduce this problem to some extent, but other correction approaches should be considered. One of the most promising strategies to achieve disambiguation is based on the preferential co-occurrence of other concepts [9;10]. For example, concepts generally co-published with PSA meaning Poultry Science Association, will be very different from concepts co-published with PSA representing prostate specific antigen. Based on these preferential co-occurring concepts, one can assign the correct meaning to an ambiguous term.

Besides disclosure, summary, and visualization of known facts using co-publication, one could also discover novel linkages among genes and between genes and other concepts. One possibility to identify unpublished, but plausible links, is to screen for black squares surrounded by red ones in a clustered co-occurrence heat map as shown in Figure 5. The fact that a particular gene-disease combination was not found in MEDLINE (black square), but clustered together with other co-published gene-disease pairs (red squares), could indicate an unpublished association. This approach shows analogies with the Swanson discovery framework in which concept A is known to relate to B and B is associated with C [49;50]. Combining all data, the deduction that A relates to C can be hypothesised and tested [49;51-53].

Conclusions

CoPub Mapper is a program that identifies and rates co-published genes and keywords starting from a single concept search or batch-wise from a set of genes. Its modularity and pre-calculated co-occurrences allow for quick and versatile querying. The regular-expression search strategy and homonym correction makes the keyword database comprehensive and less contaminated with false positive classifications. CoPub Mapper

can be used to summarize, evaluate and categorise annotated genes from microarray analyses based on co-occurrences with biological keywords and other published genes.

Availability and requirements

The CoPub Mapper program is available for free use at this URL: http://www.bioasp.nl/ or http://www.erasmusmc.nl/gatcplatform/

Authors' contributions

GJ, SvB, and JP conceived the approach and participated in the early design. BTFA and AV developed and optimised the software. TR developed and performed the homonym correction algorithm. RJ performed and interpreted the AUC ROC analyses and SV performed the MEDLINE gene and keyword searches. The project was supervised by GJ, TR and JP. All authors read and approved the final manuscript.

Acknowledgements

We thank Edwin van den Heuvel, Victor de Jager, Rene van Schaik, Jacob de Vlieg, and BioASP for their support, NLM (National Library of Medicine) for licensing of MEDLINE and Jan Kors and Jeannette Kluess for careful reading of the manuscript.

References

1. Brown PO, Botstein D: **Exploring the new world of the genome with DNA microarrays.** *Nat Genet* 1999, **21:** 33-37.

2. Duggan DJ, Bittner M, Chen Y, Meltzer P, Trent JM: **Expression profiling using cDNA microarrays.** *Nat Genet* 1999, **21:** 10-14.

3. de Bruijn B, Martin J: **Getting to the (c)ore of knowledge: mining biomedical literature.** *Int J Med Inf* 2002, **67:** 7-18.

4. Hirschman L, Park JC, Tsujii J, Wong L, Wu CH: **Accomplishments and challenges in literature data mining for biology.** *Bioinformatics* 2002, **18:** 1553-1561.

5. Mack R, Hehenberger M: **Text-based knowledge discovery: search and mining of life-sciences documents.** *Drug Discov Today* 2002, **7:** S89-S98.

6. Shatkay H, Feldman R: **Mining the biomedical literature in the genomic era: an overview.** *J Comput Biol* 2003, **10:** 821-855.

7. Pearson H: **Biology's name game.** *Nature* 2001, **411:** 631-632.

8. Wain HM, Lush MJ, Ducluzeau F, Khodiyar VK, Povey S: **Genew: the Human Gene Nomenclature Database, 2004 updates.** *Nucleic Acids Res* 2004, **32:** D255-D257.

9. Weeber M, Schijvenaars BJ, Van Mulligen EM, Mons B, Jelier R, Van Der Eijk CC, Kors JA: **Ambiguity of Human Gene Symbols in LocusLink and**

MEDLINE: Creating an Inventory and a Disambiguation Test Collection. *Proc AMIA Symp* 2003, 704-708.

10. Liu H, Johnson SB, Friedman C: **Automatic resolution of ambiguous terms based on machine learning and conceptual relations in the UMLS.** *J Am Med Inform Assoc* 2002, **9:** 621-636.

11. Chang JT, Schutze H, Altman RB: **Creating an online dictionary of abbreviations from MEDLINE.** *J Am Med Inform Assoc* 2002, **9:** 612-620.

12. Pustejovsky J, Castano J, Cochran B, Kotecki M, Morrell M: **Automatic extraction of acronym-meaning pairs from MEDLINE databases.** *Medinfo* 2001, **10:** 371-375.

13. Wren JD, Garner HR: **Heuristics for identification of acronym-definition patterns within text: towards an automated construction of comprehensive acronym-definition dictionaries.** *Methods Inf Med* 2002, **41:** 426-434.

14. Tanabe L, Wilbur WJ: **Generation of a large gene/protein lexicon by morphological pattern analysis.** *J Bioinform Comput Biol* 2004, **1:** 611-626.

15. Yeganova L, Smith L, Wilbur WJ: **Identification of related gene/protein names based on an HMM of name variations.** *Comput Biol Chem* 2004, **28:** 97-107.

16. Zhou G, Zhang J, Su J, Shen D, Tan C: **Recognizing names in biomedical texts: a machine learning approach.** *Bioinformatics* 2004, **20:** 1178-1190.

17. Yandell MD, Majoros WH: **Genomics and natural language processing.** *Nat Rev Genet* 2002, **3:** 601-610.

18. Van Der Eijk CC, Van Mulligen EM, Kors JA, Mons B, Van Den Berg J: **Constructing an associative concept space for literature-based discovery.** *J Am Soc Inf Sci Technol* 2004, **55:** 436-444.

19. Jelier R, Jenster G, Dorssers LC, Van Der Eijk CC, Van Mulligen EM, Mons B, Kors JA: **Co-occurrence based meta-analysis of scientific texts: retrieving biological relationships between genes.** *Bioinformatics* 2005, in press.

20. Swanson DR: **Medical literature as a potential source of new knowledge.** *Bull Med Libr Assoc* 1990, **78:** 29-37.

21. Chaussabel D, Sher A: **Mining microarray expression data by literature profiling.** *Genome Biol* 2002, **3:** RESEARCH 0055.

22. Becker KG, Hosack DA, Dennis G, Jr., Lempicki RA, Bright TJ, Cheadle C, Engel J: **PubMatrix: a tool for multiplex literature mining.** *BMC Bioinformatics* 2003, **4:** 61.

23. Jenssen TK, Laegreid A, Komorowski J, Hovig E: **A literature network of human genes for high-throughput analysis of gene expression.** *Nat Genet* 2001, **28:** 21-28.

24. Masys DR, Welsh JB, Lynn FJ, Gribskov M, Klacansky I, Corbeil J: **Use of keyword hierarchies to interpret gene expression patterns.** *Bioinformatics* 2001, **17:** 319-326.

25. Raychaudhuri S, Chang JT, Imam F, Altman RB: **The computational analysis of scientific literature to define and recognize gene expression clusters.** *Nucleic Acids Res* 2003, **31:** 4553-4560.

26. Hu Y, Hines LM, Weng H, Zuo D, Rivera M, Richardson A, LaBaer J: **Analysis of genomic and proteomic data using advanced literature mining.** *J Proteome Res* 2003, **2:** 405-412.

27. Glenisson P, Coessens B, Van Vooren S, Mathys J, Moreau Y, De Moor B: **TXTGate: profiling gene groups with text-based information.** *Genome Biol* 2004, **5:** R43.

28. Tanabe L, Scherf U, Smith LH, Lee JK, Hunter L, Weinstein JN: **MedMiner: an Internet text-mining tool for biomedical information, with application to gene expression profiling.** *Biotechniques* 1999, **27:** 1210-1217.

29. Chiang JH, Yu HC, Hsu HJ: **GIS: a biomedical text-mining system for gene information discovery.** *Bioinformatics* 2004, **20:** 120-121.

30. Lin SM, McConnell P, Johnson KF, Shoemaker J: **MedlineR: an open source library in R for Medline literature data mining.** *Bioinformatics* 2004, **20:** 3659-3661.

31. Stapley BJ, Benoit G: **Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in Medline abstracts.** *Pac Symp Biocomput* 2000, 529-540.

32. Iliopoulos I, Enright AJ, Ouzounis CA: **Textquest: document clustering of Medline abstracts for concept discovery in molecular biology.** *Pac Symp Biocomput* 2001, 384-395.

33. Raychaudhuri S, Schutze H, Altman RB: **Using text analysis to identify functionally coherent gene groups.** *Genome Res* 2002, **12:** 1582-1590.

34. Liu G, Loraine AE, Shigeta R, Cline M, Cheng J, Valmeekam V, Sun S, Kulp D, Siani-Rose MA: **NetAffx: Affymetrix probesets and annotations.** *Nucleic Acids Res* 2003, **31:** 82-86.

35. **The Online Plain Text English Dictionary** [http://msowww.anu.edu.au/~ralph/OPTED]

36. Zeeberg BR, Riss J, Kane DW, Bussey KJ, Uchio E, Linehan WM, Barrett JC, Weinstein JN: **Mistaken Identifiers: Gene name errors can be introduced inadvertently when using Excel in bioinformatics.** *BMC Bioinformatics* 2004, **5:** 80.

37. **National Library of Medicine** [http://www.nlm.nih.gov/]

38. **Human and Animal Cell Line Names** [http://www.biotech.ist.unige.it/cldb/cname-1c.html]

39. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns.** *Proc Natl Acad Sci U S A* 1998, **95:** 14863-14868.

40. **European Bioinformatics Institute** [http://srs.ebi.ac.uk/]

41. **Gene Ontology** [http://www.geneontology.org/]

42. **BioCarta** [http://www.biocarta.com/]

43. Smid M, Dorssers LC, Jenster G: **Venn Mapping: clustering of heterologous microarray data based on the number of co-occurring differentially expressed genes.** *Bioinformatics* 2003, **19:** 2065-2071.

44. Hanley JA, McNeil BJ: **The meaning and use of the area under a receiver operating characteristic (ROC) curve.** *Radiology* 1982, **143:** 29-36.

45. Jansen E, Laven JS, Dommerholt HB, Polman J, Van Rijt C, Van Den HC, Westland J, Mosselman S, Fauser BC: **Abnormal gene expression profiles in human ovaries from polycystic ovary syndrome patients.** *Mol Endocrinol* 2004, **18:** 3050-3063.

46. Guzick DS: **Polycystic ovary syndrome.** *Obstet Gynecol* 2004, **103:** 181-193.

47. Solomon CG: **The epidemiology of polycystic ovary syndrome. Prevalence and associated disease risks.** *Endocrinol Metab Clin North Am* 1999, **28:** 247-263.

48. Zdobnov EM, Lopez R, Apweiler R, Etzold T: **The EBI SRS server--recent developments.** *Bioinformatics* 2002, **18:** 368-373.

49. Smalheiser NR, Swanson DR: **Using ARROWSMITH: a computer-assisted approach to formulating and assessing scientific hypotheses.** *Comput Methods Programs Biomed* 1998, **57:** 149-153.

50. Swanson DR: **Fish oil, Raynaud's syndrome, and undiscovered public knowledge.** *Perspect Biol Med* 1986, **30:** 7-18.

51. Srinivasan P, Libbus B: **Mining MEDLINE for implicit links between dietary substances and diseases.** *Bioinformatics* 2004, **20 Suppl 1:** I290-I296.

52. Weeber M, Vos R, Klein H, De Jong-Van Den Berg LT, Aronson AR, Molema G: **Generating hypotheses by discovering implicit associations in the literature: a case report of a search for new potential therapeutic uses for thalidomide.** *J Am Med Inform Assoc* 2003, **10:** 252-259.

53. Wren JD, Bekeredjian R, Stewart JA, Shohet RV, Garner HR: **Knowledge discovery by automated identification and ranking of implicit relationships.** *Bioinformatics* 2004, **20:** 389-398.

54. **Affymetrix** [http://www.affymetrix.com]

55. **HUGO Gene Nomenclature Committee** [http://www.gene.ucl.ac.uk/nomenclature/]

56. **Karolinska Institiute Alphabetic List of Specific Diseases/Disorders** [http://www.mic.ki.se/Diseases/Alphalist.html]

57. **Medical Subject Headings** [http://www.nlm.nih.gov/mesh/meshhome.html]

Chapter 4

A taxonomy-based disambiguation approach for gene names and symbols

Blaise T. F. Alako[1,2], Pieter Neerincx[1] and Jack A. M. Leunissen[1]
[1]Laboratory of Bioinformatics, Wageningen University and Research Centre, PO Box 8128, 6700 ET Wageningen, the Netherlands,
[2]Centre for BioSystems Genomics, PO Box 98, 6700 AB Wageningen, the Netherlands
blaise.alako@wur.nl, Pieter.neerincx@wur.nl, jack.leunissen@wur.nl

In preparation for submission

Abstract

Background

One of the challenges text-mining is facing is the gene nomenclature ambiguity. A gene symbol can have multiple alternative names (synonyms) and/or multiple different functional assignments (homonyms) related to its biological function, its physiological location, or even as the result of researchers' conflicts of interest. Moreover, this lack of a controlled vocabulary in gene nomenclature is complicated even more by common problems such as alternative spellings and literal spelling mistakes. This ambiguity makes text mining, information retrieval, information extraction and data mining in the biological field a difficult task.

Results

We have developed a disambiguation methodology that relies on the hypothesis that if we can categorize ambiguous gene symbols correctly in terms of their biological function(s), then the taxonomic relationship of genes' species in each category can be used to tag that category. Unlike previous species-specific approaches, our disambiguation approach is able to cover the entire taxonomic spectrum from viruses, prokaryotes and archaea to eukaryotes. We use the NCBI taxonomy database to resolve all ambiguous gene symbols in the UniProt Knowledgebase (UniProtKB) with an overall 94 % precision and 82 % recall.

Conclusion

Our algorithm uses a naïve Bayes classifier to solve problems such as: "Given an ambiguous gene symbol and a species name what is its most likely functional assignment?" or "What biological terms and gene symbol synonyms can be used in a Boolean query to efficiently and unambiguously search the literature for documents pertaining to a gene of interest?"

Background

Molecular biology is undoubtedly one of the more recent scientific fields that has moved from technology and technique-based research to information driven research (1). This is emphasized by the wealth of information and data from sequencing projects and exponential publications about genes and groups of genes. Swanson's work on finding implicit relations between facts in publications promoted a greater interest in using literature as a knowledge base for hypothesis generation (2-6). Since then several informatics tools for literature mining have been developed for extracting information based on term co-occurrence such as CoPub Mapper and many more (7-11). In order to retrieve sound facts and obtain knowledge it is crucial to pose effective and proper questions. However, at present mining the literature is complicated by the presence of ambiguity that is inherent to natural language in a publication's abstract or full paper. Naturally, this leads to the question: How effective are currently developed text mining tools?

The success of the Critical Assessment of Protein Structure Prediction (CASP) (12), Critical Assessment of Prediction of Interaction (CAPRI) (13), Critical Assessment of Micro-array Data Analysis (CAMDA) (14) and the Genome Annotation Assessment Project (GASP) (15) initiated various evaluation projects of text mining in the biological field. These are in order of their initiation: Knowledge discovery and Data mining (KDD), Text Retrieval Conference (TREC), Critical Assessment of Information Extraction in Biology (BioCreAtIvE), and Natural Language Processing for Biology (BioNLP). TREC focuses on document retrieval and classification tasks for genomics (16,17), whereas BioNLP tags the biological name in Medline abstracts and BioCreAtIvE focuses on gene mention identification and normalization and functional annotation of genes using GO (18-22).

One of the drawbacks of these evaluation approaches is their variation in performance, for example in BioCreAtIvE's "gene or protein mention" and "normalized gene list" task, a variation was found between organisms with top F-scores higher than 0.90 for yeast and 0.80 for both fly and mouse. Detailed analysis revealed that the difference between organisms could be explained at the gene nomenclature level due to extensive ambiguity in gene names, overlap of gene names with English term, complex multiword gene names, and difficulty in associating ambiguous names with the correct gene identifier (21,22). Additionally, a substantial ambiguity in gene nomenclature has been shown within and across eukaryote species, with English terms and with medical terms (23).

So far, efforts have mainly addressed the quantification of ambiguity restricted to specific organisms. To our knowledge the problem of ambiguity quantification to date has been addressed mainly in 21 model organisms with various solutions proposed to reduce name ambiguity (23). The main solutions have been thesaurus-based (24), using classifiers (25,26) and context based-coupled classifiers disambiguation (27-30). Schuemie and co-workers (31) give an overview of the word sense disambiguation in the biomedical domain.

Ambiguity and the disambiguation process are highly complex, particularly if we consider the entire taxonomic spectrum, as is the case in this paper. For example, the gene symbol CAT2 is thought to be ambiguous in rats (a category of permeases and a category of ion transfer), but also denotes a gene in fungi (yeast, baker's yeast), plants (tomato, potato, radish) and bacteria (*Halobacterium salinarium, Salinibacter ruber*). The functions of gene CAT2 can be generalized as permeases and catalases in bacteria and plants, respectively as can be taken from ***Figure 1-a,b*** generated with TreeDomViewer (32).

TreeDomViewer displays the structural domains shared by the different sequences annotated as CAT2.

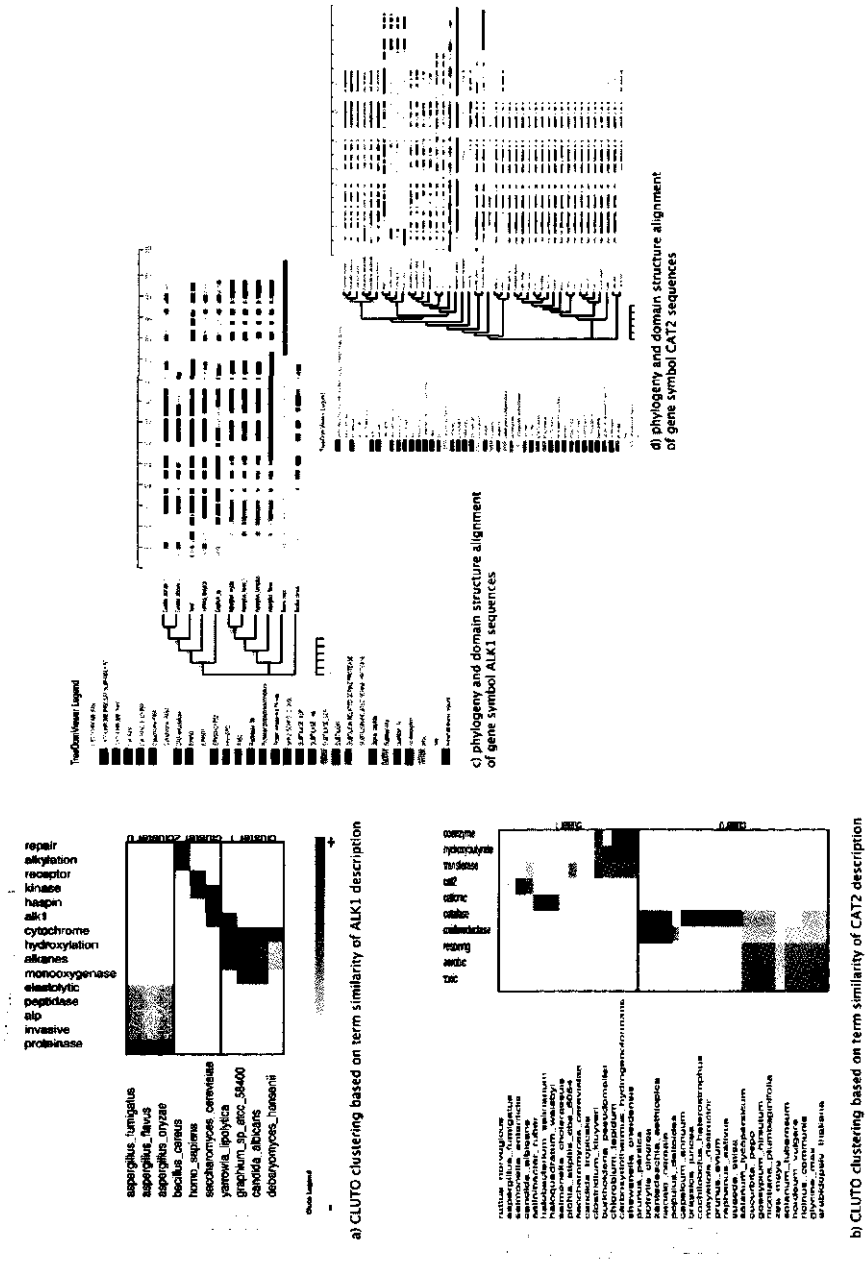
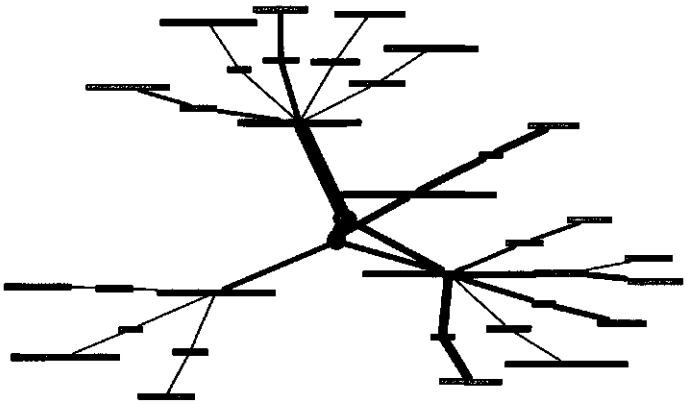
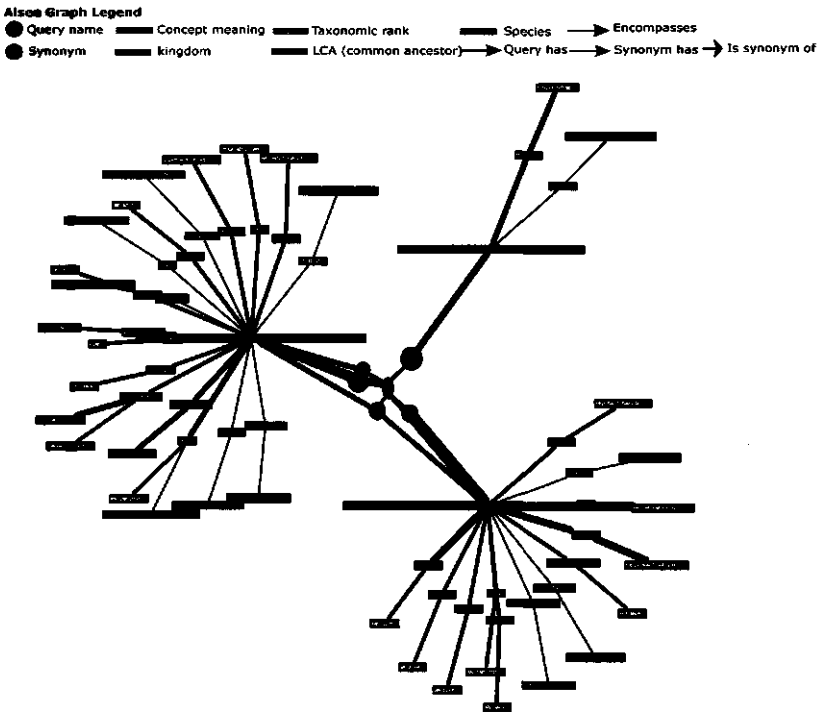


Figure 1: Phenotype (text data) (a,b) and genotype (Sequence data) (c,d) correlation of 2 gene symbols, ALK1, CAT2.

Another example of an ambiguous gene symbol used for illustration in this research is ALK1, which refers to a gene found in fungi and bacteria. ***Figure 1-c,d*** shows the existence of two distinct groups in fungi, described respectively as cytochrome P450 and permease-peptidase related, whereas in bacteria ALK1 refers to a helix-hairpin-helix superfamily-based excision DNA repair protein. These two examples illustrate the ambiguity within species (intra-species) as well as between species (inter-species).
Furthermore, CAT2 also displays the problem of synonymy with CAT, CAT1, ROP and SU2, encompassing species from a broad taxonomic spectrum, namely archaea, bacteria, fungi, metazoa and plants (***Figure 2-b***). Some of these synonyms share the same biological category, that is, function, as is the case with CAT and CAT2, whereas SU2 only shares one out of two CAT2 categories. This shows that the ambiguity problem becomes highly complex if we do consider gene symbols and their synonyms alike. This observed gene symbol ambiguity might be explained amongst others by variation on symbol's spelling and researchers' conflict of interest (33-35), occurrence of multiple names for the same genes (synonyms) and gene symbols referring to different biological function (homology).
The ambiguity issue that is addressed in this paper will focus on gene symbol homonymy and gene symbol synonymy. Additionally we intend to extend the definition of gene ambiguity beyond non-uniqueness in a database, stop words, general English words or non-biological terms to a gene having multiple concepts (biological functions).
First we will quantify the ambiguity problem and subsequently assign the biological meaning or function to each gene symbol (discrimination task). Once plausible senses are obtained for each gene symbol, we will assign the proper function to a symbol given a species using the disambiguation task. Therefore, our approach should be viewed as species taxonomy-pivoted gene function categorization.
Taxonomy may refer to relationship schemes other than hierarchies such as network structure as well as being a simple organization of objects into groups, or even an alphabetical list. In this paper we use taxonomy in the sense of an expert evolutionary classification of organisms.
Our approach is based on the fact that most assignments of annotation for molecular function rely, at least partially, on the assumption that genes with similar sequences also display similar biological functions. This implies that sequences are evolutionary related to a certain extent, the relationship denoting a common latest ancestor prior to the speciation event. Thus species of a certain taxonomic rank have the same biological function for a specific gene.
Provided that we can discriminate properly between gene symbols in terms of their function, then the latest common species ancestor of a gene symbol within a category can be used to tag that category. For example, if the gene symbol ALK1 has shown 3 distinct biological functions during the discrimination process, we assume that each group is monophyletic i.e. of one race, if not transformed to monophyletic, and the latest common ancestor (LCA) of each group is used to assign a function to that category in the disambiguation process.
In the following sections we will present the results and evaluation of the different subtasks namely gene symbol's sense discrimination and gene symbol's sense disambiguation used in our methodology.



a) Taxonomic-based disambiguation of ALK1. Species having aspergillus as the latest common ancestor have the same biological function as shown the graph.



b) Taxonomic-based disambiguation of CAT2. Species having magnoliophyta as the latest common ancestor have the same biological function as shown the graph.

Figure 2 – Taxonomic based disambiguation of two gene symbols ALK1(a) and CAT2(b).

Results Evaluation and Discussion

Gene symbol's sense discrimination and quantification of ambiguity

Gene symbol senses discrimination is the task of finding the potential biological function related to a gene symbol. Each sense is further labeled with an unambiguous meaning. This stage can also be used to automatically assist in the construction of ontologies that is currently done manually. Throughout this work, we address gene symbol or short form rather than its long form (expanded). Occurrence of the gene name in more than one UniProtKB entry was the starting point for our ambiguity quantification, but often these symbols converged to the same biological function. Thus, our initial data set comprised 243,366 gene symbols occurring more than once in UniProtKB and was minimized to 10, 865 ambiguous symbols, addressing more than one actual biological function. After obtaining the plausible biological function related to each ambiguous gene symbol during the discrimination stage (***see Materials and Methods***), we examined the gene symbol's senses distribution. As shown in ***Figure 3-a,b*** we identified 57% of the total 243,366 putative ambiguous gene symbols as displaying the same biological function, 40% with 2 distinct biological functions, 3% with 3 and the remaining genes with 4 to 8 biological functions. We further investigated the distribution of the 10,865 symbols showing more than one biological function among SwissProt taxonomic classes to gain an overview of the classes with more confusing names. Our main finding was that 21% occurred in bacteria, 21% in human and 18% in rodents as shown in ***Figure 3-c.***

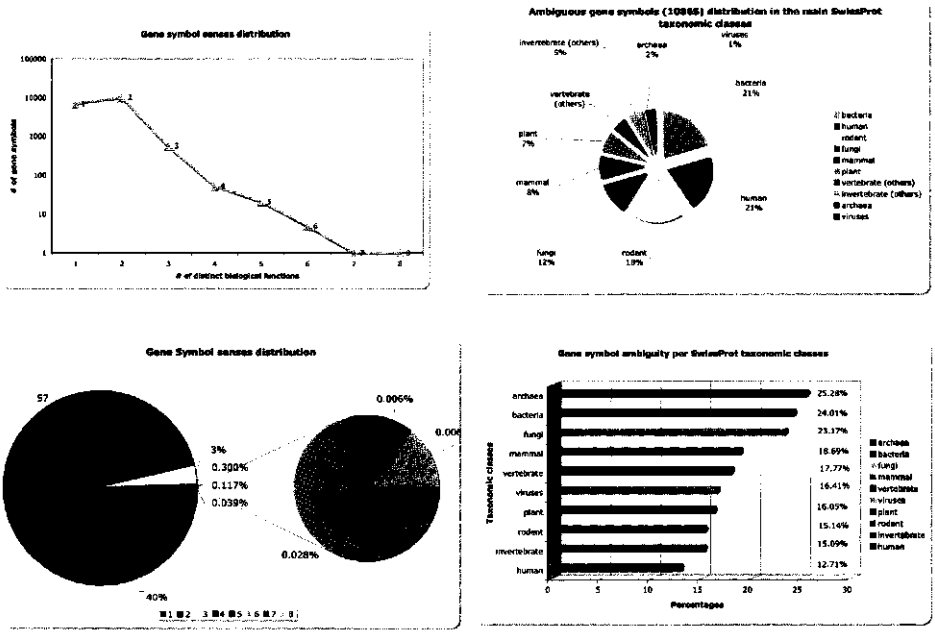


Figure 3 – Quantification of ambiguous gene symbols in UniProtKB.

However, this figure also showed an overlap in gene symbols between taxonomic classes, which will be illustrated further on. This overlap could be among others assigned to ambiguity introduced via homology inference between model species and organisms of interest.

Subsequently we sought to quantify gene symbols shared by different taxonomic classes using Venn Master (36) to analyze the 10,865 ambiguous gene symbols. ***Figure 4*** summarizes our findings: mammals, plants, vertebrates, rodents, fungi and humans share 49 gene symbols whereas mammals, bacteria, plants vertebrates, rodents and fungi share only 11 gene symbols. Furthermore, bacteria, plants and human share 99 gene symbols.

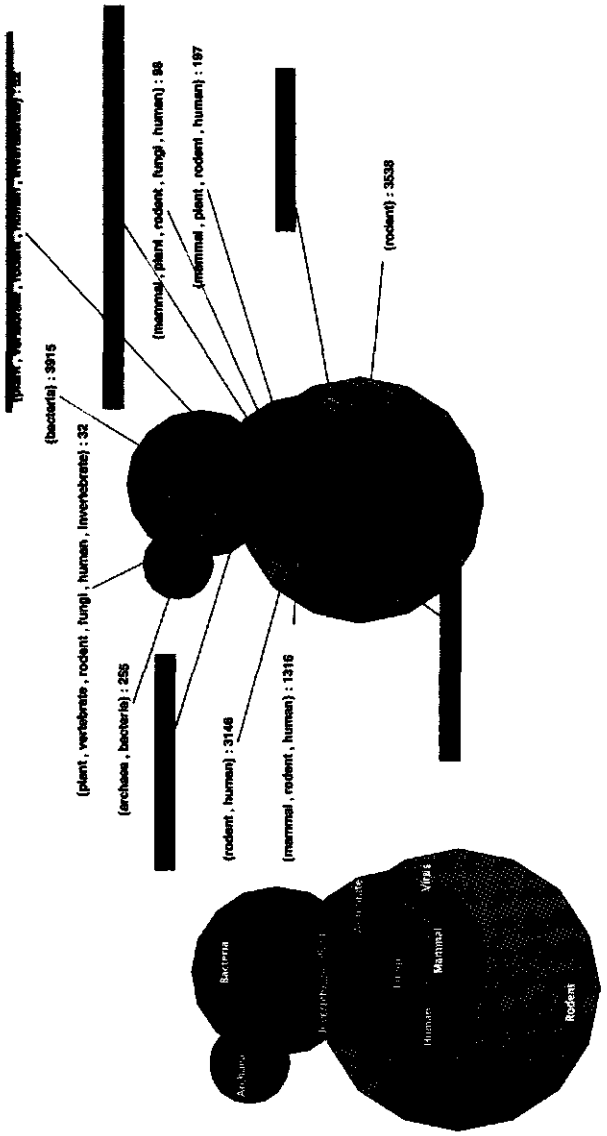


Figure 4- Ambiguous gene symbols (senses >=2) distribution across main SwissProt taxonomic classes i.e. rodent, human, archaea, plant, vertebrate (others) invertebrates (others), mammal, fungi, bacteria, viruses.

Inaccuracy in gene nomenclature by scientists, conflict of interest or pure coincidence of gene symbols having multiple alternative names, may account for our observations. Having observed the interspecies gene nomenclature ambiguity lead us to the question in which taxonomic class is the ambiguity more prevalent? To address this question, we quantified the ambiguity within the main SwissProt taxonomic classes and found that ambiguous gene symbols accounted for 362 out of 1,432 in archaea (25.28%), in bacteria 3915 out of 16,303 (24.01%) and in fungi 2,227 out of its 9610 (23.17%) gene symbol entries.(***Figure 3-d***). Our findings can be attributed to the fact that archaeal, bacterial and fungal taxonomic classes have more species and gene symbol entries than any other class in the UniProtKB. Entries for human showed the least ambiguous gene symbols, namely 3,824 out of its 30,090 (12.71%) gene symbols; this can be attributed to the international effort in improving and standardizing human genes. Taken together, these results strongly show the extent and spread of gene nomenclature ambiguity within species and between species.

Hereafter we evaluated our methodology at the discrimination and the disambiguation task level, respectively, to determine its suitability at detecting ambiguous gene symbols. In order to address this question, we evaluated the correlation between the phenotype (textual information) and genotype (sequence information). We randomly selected 100 ambiguous gene symbols out of the 10,865 available. For each randomly selected symbol we extracted the corresponding protein sequence from the UniProtKB. We used TreeDomViewer for displaying the phylogeny with structural domain information supporting each clade and CLUTO to cluster textual (phenotype) information extracted from UniProtKB, related to the same gene symbol (***Figure 1***). Then we manually compared the two generated figures; in 86 % of the cases the phenotype analysis correlated well with the genotype analysis ***(see additional files)***. For 10% of the cases the respective domain could not be predicted for the sequence, therefore making it difficult to support the members of each clade based on their structural domain(s). In 4% of the cases we predicted more groups than actually was the case. A closer look at the feature space revealed that some gene symbol entries were solely described by the gene symbols itself without any other supplementary information. Moreover, some symbols were sparsely annotated and/or with very generic terms such as "hypothetical protein". Such general descriptive terms were excluded from the analysis (***see Materials and Methods***). The correction of the above accounted for a total of 90% correlation between the phenotype and the genotype comparison.

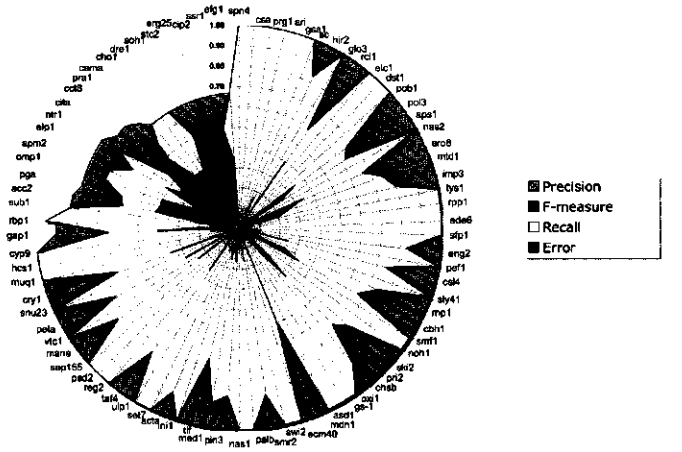
Gene symbol's sense disambiguation

In order to address the experimental analysis of our method, we opted for the leave-one-out cross validation given the sparse data at hand. Again we randomly selected 100 gene symbols from the ambiguous set. For each gene symbol, we trained the Naïve Bayes classifier on n-1 instances ***(see additional files)***. These instances were the latest common ancestor (LCA) of the species within a biological function class and the corresponding biological function. We then iterated the classification in such a way that all species instances were used as a testing set and the average performance was reported as the microaveraging and macroaveraging metric of our classifier. With microaveraging we achieved 93.52 % precision, 82.12 % recall and 85.57 % F-measure on average and with macroaveraging metric we reported 82.81% precision, 85.97 % recall and 82.81% F-measure. An excerpt of the performance metrics on 20 ambiguous genes symbols out of the 100 randomly selected from the total 10, 865 ambiguous gene symbols is reported in ***Table 1 and Figure 5***.

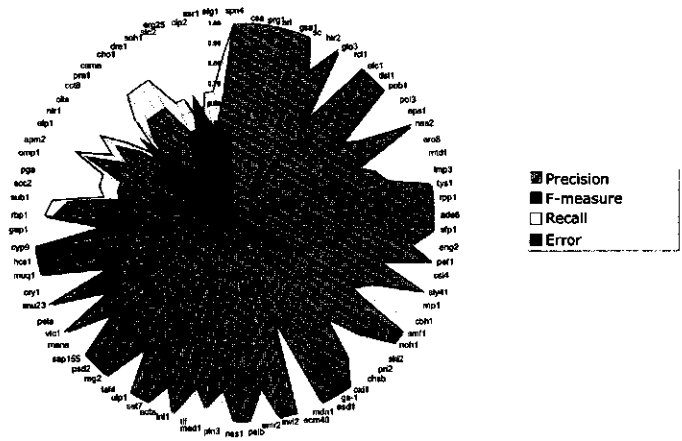
symbols	micro_P	micro_R	micro_F1	micro_E	micro_A	macro_P	macro_R	macro_F1	macro_E	macro_A
spn4	1	1	1	0	1	1	1	1	0	1
cas	1	1	1	0	1	1	1	1	0	1
prg1	1	1	1	0	1	1	1	1	0	1
erf	1	1	1	0	1	1	1	1	0	1
gas1	1	1	1	0	1	1	1	1	0	1
ac	1	1	1	0	1	1	1	1	0	1
hir2	1	0.8333	0.889	0.1667	0.8333	0.8333	0.8333	0.8333	0.1667	0.8333
gle3	1	1	1	0	1	1	1	1	0	1
rd1	1	0.6	0.7336	0.4	0.6	0.6	0.6	0.6	0.4	0.6
sic1	1	1	1	0	1	1	1	1	0	1
dst1	1	1	1	0	1	1	1	1	0	1
pob1	1	1	1	0	1	1	1	1	0	1
pol3	1	0.6667	0.778	0.3333	0.6667	0.6667	0.6667	0.6667	0.3333	0.6667
sps1	1	0.625	0.7503	0.2498	0.7503	0.7503	0.7503	0.7503	0.2498	0.7503
nam2	1	1	1	0	1	1	1	1	0	1
ero8	1	0.6667	0.778	0.3333	0.6667	0.6667	0.6667	0.6667	0.3333	0.6667
mbd1	1	0.75	0.8335	0.25	0.75	0.75	0.75	0.75	0.25	0.75
imp3	1	0.8	0.8668	0.2	0.8	0.8	0.8	0.8	0.2	0.8
tys1	1	1	1	0	1	1	1	1	0	1
rpp1	1	1	1	0	1	1	1	1	0	1
adh6	1	1	1	0	1	1	1	1	0	1
efp1	1	1	1	0	1	1	1	1	0	1
cog2	1	0.8333	0.889	0.111	0.889	0.889	0.889	0.889	0.111	0.889
pef1	1	1	1	0	1	1	1	1	0	1
cst4	1	0.75	0.8335	0.25	0.75	0.75	0.75	0.75	0.25	0.75
shy41	1	1	1	0	1	1	1	1	0	1
rnp1	1	0.75	0.8335	0.25	0.75	0.75	0.75	0.75	0.25	0.75
clh1	1	0.6667	0.778	0.3333	0.6667	0.6667	0.6667	0.6667	0.3333	0.6667
swf1	1	1	1	0	1	1	1	1	0	1
noh1	1	1	1	0	1	1	1	1	0	1
sld2	1	0.7222	0.8057	0.1667	0.8333	0.8333	0.8333	0.8333	0.1667	0.8333
prt2	1	0.6667	0.778	0.3333	0.6667	0.6667	0.6667	0.6667	0.3333	0.6667
chsb	1	0.875	0.9167	0.125	0.875	0.875	0.875	0.875	0.125	0.875
cid1	1	1	1	0	1	1	1	1	0	1
ge-1	1	1	1	0	1	1	1	1	0	1
add1	1	1	1	0	1	1	1	1	0	1
mdn1	1	0.5	0.667	0.5	0.5	0.5	0.5	0.5	0.5	0.5
ecm40	1	1	1	0	1	1	1	1	0	1
swi2	1	1	1	0	1	1	1	1	0	1
smr2	1	0.8333	0.889	0.1667	0.8333	0.8333	0.8333	0.8333	0.1667	0.8333
pab	1	1	1	0	1	1	1	1	0	1
nas1	1	1	1	0	1	1	1	1	0	1
pin3	1	0.8333	0.889	0.1667	0.8333	0.8333	0.8333	0.8333	0.1667	0.8333
med1	1	0.875	0.9167	0.0625	0.9375	0.9375	0.9375	0.9375	0.0625	0.9375
tif	1	0.6666	0.7668	0.16	0.84	0.84	0.84	0.84	0.16	0.84
lrd1	1	1	1	0	1	1	1	1	0	1
acta	1	0.875	0.9167	0.125	0.875	0.875	0.875	0.875	0.125	0.875
set7	1	1	1	0	1	1	1	1	0	1
uip1	1	1	1	0	1	1	1	1	0	1
taf4	1	0.75	0.8335	0.25	0.75	0.75	0.75	0.75	0.25	0.75
reg2	1	1	1	0	1	1	1	1	0	1
pad2	1	1	1	0	1	1	1	1	0	1
sap155	1	1	1	0	1	1	1	1	0	1
mana	1	0.75	0.8335	0.25	0.75	0.75	0.75	0.75	0.25	0.75
vtc1	1	1	1	0	1	1	1	1	0	1
pela	1	0.7	0.8002	0.3	0.7	0.7	0.7	0.7	0.3	0.7
snu23	1	1	1	0	1	1	1	1	0	1
cry1	1	0.6429	0.7621	0.3571	0.6429	0.6429	0.6429	0.6429	0.3571	0.6429
muq1	1	1	1	0	1	1	1	1	0	1
hca1	1	1	1	0	1	1	1	1	0	1
cyp9	1	1	1	0	1	1	1	1	0	1
gap1	0.9	0.65	0.6669	0.4	0.6	0.6	0.65	0.6	0.4	0.6
rtp1	0.9	0.95	0.8667	0.1	0.9	0.9	0.95	0.9	0.1	0.9
aub1	0.8333	0.8333	0.5333	0.1667	0.8333	0.8333	0.9167	0.8333	0.1667	0.8333
ecc2	0.8333	0.5833	0.6668	0.4167	0.5833	0.5833	0.6667	0.5833	0.4167	0.5833
pga	0.8	0.6	0.6668	0.4	0.6	0.6	0.7	0.6	0.4	0.6
omp1	0.8	0.7	0.6002	0.4	0.6	0.6	0.7	0.6	0.4	0.6
apm2	0.8	0.7	0.7334	0.2	0.8	0.8	0.8668	0.8	0.2	0.8
elp1	0.8	0.6	0.6668	0.4	0.6	0.6	0.7	0.6	0.4	0.6
ntr1	0.8	0.7	0.7334	0.3	0.7	0.7	0.8	0.7	0.3	0.7
clb	0.75	0.4582	0.5417	0.4168	0.5833	0.5833	0.6667	0.5833	0.4168	0.5833
cct8	0.75	0.375	0.5003	0.625	0.375	0.375	0.5	0.375	0.625	0.375
pre1	0.7143	0.5714	0.5191	0.4286	0.5714	0.5714	0.7143	0.5714	0.4286	0.5714
cam4	0.6667	0.6667	0.6667	0.3333	0.6667	0.6667	0.8333	0.6667	0.3333	0.6667
cho1	0.6667	0.6667	0.6667	0.3333	0.6667	0.6667	0.8333	0.6667	0.3333	0.6667
dre1	0.6667	0.6667	0.6667	0.3333	0.6667	0.6667	0.8333	0.6667	0.3333	0.6667
soh1	0.6667	0.5	0.5557	0.5	0.5	0.5	0.6667	0.5	0.5	0.5
stc2	0.6667	0.5	0.5557	0.5	0.5	0.5	0.6667	0.5	0.5	0.5
arg25	0.6667	0.3333	0.4447	0.6667	0.3333	0.3333	0.5	0.3333	0.6667	0.3333
clp2	0.6667	0.5	0.5557	0.5	0.5	0.5	0.6667	0.5	0.5	0.5
ser1	0.6667	0.5	0.5557	0.5	0.5	0.5	0.6667	0.5	0.5	0.5
efg1	0.6667	0.6667	0.6667	0.3333	0.6667	0.6667	0.8333	0.6667	0.3333	0.6667
nudg	1	0.6667	0.778	0.3333	0.6667	0.6667	0.6667	0.6667	0.3333	0.6667
sps1	1	1	1	0	1	1	1	1	0	1
pas1	1	1	1	0	1	1	1	1	0	1
pox6	1	0.6	0.7336	0.4	0.6	0.6	0.6	0.6	0.4	0.6
cid1	1	1	1	0	1	1	1	1	0	1
idh2	1	0.75	0.8335	0.25	0.75	0.75	0.75	0.75	0.25	0.75
pih4	1	0.875	0.9167	0.125	0.875	0.875	0.875	0.875	0.125	0.875
rpa3	1	1	1	0	1	1	1	1	0	1
opi3	1	1	1	0	1	1	1	1	0	1
pup2	1	1	1	0	1	1	1	1	0	1
ntr2	1	0.7777	0.8333	0.2223	0.7777	0.7777	0.7777	0.7777	0.2223	0.7777
air1	1	1	1	0	1	1	1	1	0	1
ser3	1	1	1	0	1	1	1	1	0	1
cdc54	1	1	1	0	1	1	1	1	0	1
nat4	1	1	1	0	1	1	1	1	0	1
taf9	1	0.8	0.8668	0.2	0.8	0.8	0.8	0.8	0.2	0.8
agaa	1	0.5	0.667	0.5	0.5	0.5	0.5	0.5	0.5	0.5

Table 1: Performance metric on 100 ambiguous gene symbols randomly selected.

The significant difference between microaveraging and macroaveraging could be attributed to the fact that the different categories show a very different generality i.e. the percentage of species and / or LCA (positive training) that belong to that category.
Thus the ability of the classifier to behave well also on categories with low generality will be emphasized by macroaveraging and much less so by microaveraging.



a) Micro performance metric.



b) Macro performance metric.

Figure 5 – Micro (a) and Macro (b) averaging performance metric of GeneIlluminator via Leave-one-out crossvalidation on the 100 randomly selected gene symbols and corresponding species.

The association of clustering and taxonomy-based tagging

Throughout our methodology we have combined clustering and species taxonomic tagging. Clustering and taxonomic tagging are similar when new annotations are introduced in the UniProtKB databases, i.e. they have to be run again. The quality of the clustering technique depends heavily on the feature space, while species taxonomy tagging relies on the existence of a taxonomic description and entry of a given species in the NCBI taxonomy database.

The above analysis shows that a prior selection of a comprehensive database such as UniProtKB and the NCBI taxonomic database coupled to a post-retrieval clustering followed by taxonomic rank tagging offers quality, maintenance, flexibility and performance benefits. This is an important advantage over thesaurus-based, rule-based disambiguation and previous gene symbol sense disambiguation approaches.

Future work

We plan to develop a web interface to navigate and resolve the ambiguity through other classifier besides the current Naïve Bayes. An area under curve (AUC) would be a sound statistical measure to select for the evaluation of classifier quality. The latter interface should also help recovering literature and other information pertaining to a gene of interest, for example disambiguating PubMed abstracts.

Conclusions

We present a disambiguation algorithm that integrates clustering, discriminating, tagging, mapping and categorization algorithms. The methodology relies on the assumption that once function is properly categorized for an ambiguous gene name, we can uniquely tag all members of each category to a taxonomic rank, and thus map the taxonomic or species level to one of the concepts, i.e. the gene name function. This approach disambiguates gene names in a wide taxonomic spectrum, namely viruses, prokaryotes, archaea and eukaryotes, and resolves ambiguity within species as well as between species. The generated data can be used to train a Naïve Bayes classifier which can properly categorize new gene symbols given the fact that the latter is ambiguous or not and that the species is known.

This methodology attempts to answer questions such as "Which gene do you address?" in an association study or text-mining tool. It also highlights the multiple aspects (biological function) of a gene symbol. Our algorithm relies on the taxonomy (LCA – last common ancestor) as a substitute for the biological function of a gene symbol if in a context the species is known and the function is problematic. Moreover, for database curators this tool provides a good overview of the extent of ambiguity in their repository, as well as an efficient way to resolve the ambiguity.

Materials and Methods

Data collection

Data were extracted from the UniProtKB (Universal Protein Knowledge based) database release 12.1 (37), which is a central hub for the collection of functional information on protein, with accurate, consistent and rich information and an accepted biological ontology, taxonomic classification and cross-referenced information. ***Figure 6*** "step 1" highlights the data collection stage. UniProtKB is made up of SwissProt (manually curated set) and TrEMBL (automatically generated).

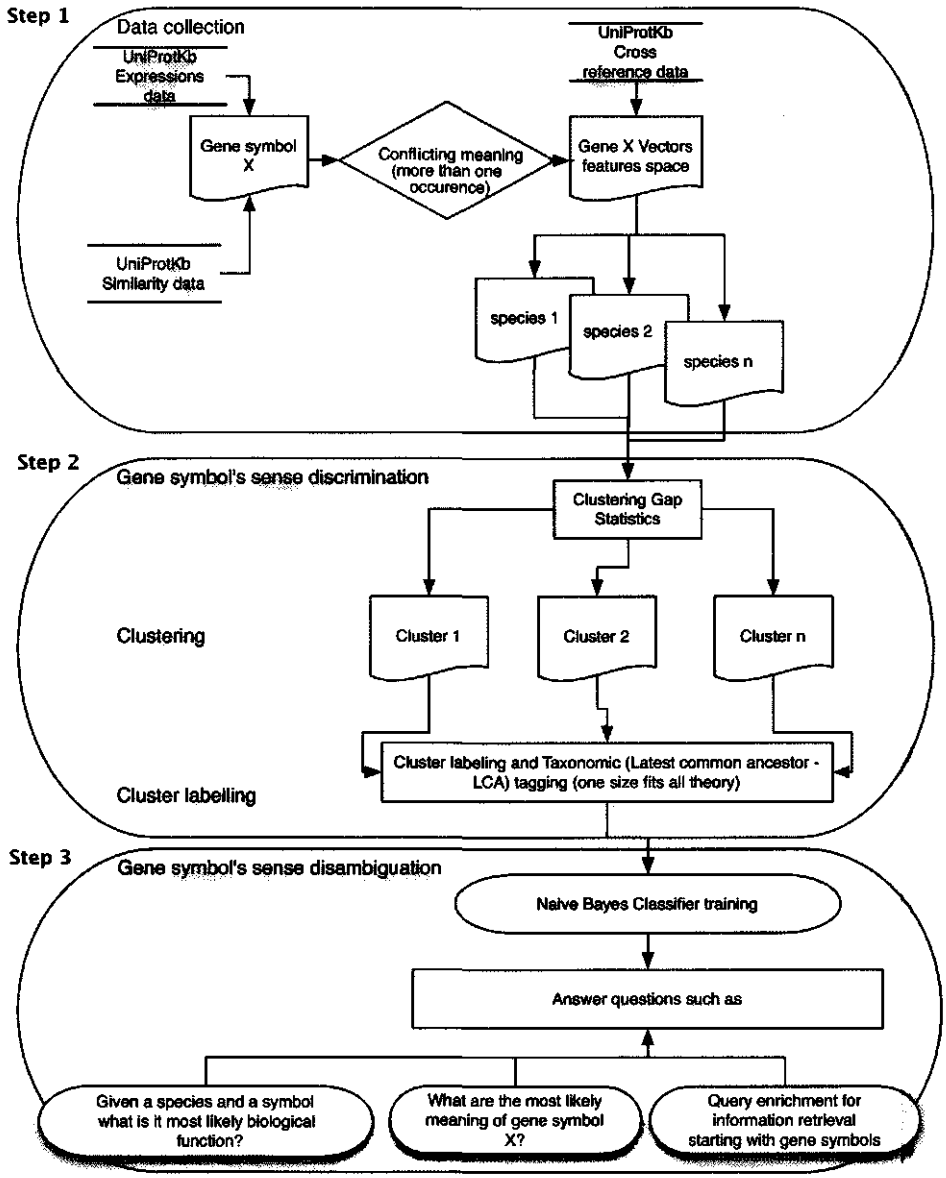


Figure 6 – Disambiguation algorithm data collection, discrimination and disambiguation workflow.

In order to investigate which name variant (primary name, alternative name) yielded greater ambiguity, we investigated the ambiguity level considering only the primary gene name in the UniProtKB. This explored the extent of the problem if scientists only referred to primary nomenclature in their research. A Perl module referred to as ***"neverExpand"*** implemented this scenario. We further measured the ambiguity extent considering the primary name and its UniProtKB alternative name(s), provided the latter was not a primary name elsewhere. This approach sought to highlight which of the nomenclature categories were mostly used and which might aggravate the ambiguity level; the ***"CondExpand"*** Perl module handled this case. Finally, the module ***"alwaysExpand"*** assumed that primary name and alternative names are likely to contribute to ambiguity. This analysis resulted in 10,865 ambiguous symbols out of 243,366 UniProtKB gene symbols occurring more than once, which was 4.5 % of the total set. ***Figure 3*** illustrates the extent of gene ambiguity in UniProtKB when using the above-mentioned approaches. The rationale in our investigation so far was to optimally quantify the ambiguity and as such, ***"AlwaysExpand"*** was adopted for our study.

Data curation:

Feature stemming

For biological terms stemming can be considered done at two levels, namely the suffixes and the prefixes. Taking the prefix case, deacetylation stemmed to acetylation conveys a complete opposite meaning. To further illustrate this, we consider the following processes or actions: maturation, differentiation, and inhibition. Yet once maturation is stemmed to mature (fully developed), the biological meaning is changed. Furthermore, enzymes are suffixed with –ase (e.g. peptidase), and stemming carried out on disease (which of course is not an enzyme) results in "dise", which is neither a biological term nor an English term, but a French derivate of the verb "dire" (to say). Hence, using stemming in vector feature construction in this biological context might lead to a serious semantic error and after judging the risk stemming brought into play, we discarded it.

Stop list generation and biological term selection

The appropriate content identifiers were extracted from UniProtKB. We focused on three limited annotation fields that were likely to be highly enriched. These were the UniProtKb description field (DE), the comment field (CC) and Cross-reference fields (DR). In the next section we will explain our strategy for selecting gene symbol's content as well as the terms to be discarded from the whole set, the so-called stop word list.

Most automatic indexing efforts start with the observation that the frequency of occurrence of individual word types (that is, of distinct words) in natural language texts (biological text) has something to do with the importance of these words for the purpose of contents representation. It has been observed that words occur in natural language unevenly. Consequently, classes of words are distinguishable by their frequency of occurrence. In fact, it is known that when distinct words in a text body are arranged in a decreasing order of their frequency of occurrence, the constant rank-frequency law of Zipf can characterize the occurrence characteristics of the vocabulary,

$$frequency \times rank \sim= constant,$$

i.e. the frequency of a given word multiplied by the rank order of that word will be approximately equal to the frequency of another word multiplied by its rank (38). The law has been explained by citing a general "principle of least effort", which makes it easier for a speaker or writer to repeat certain words instead of coining new and different words, with the exception for those used in poetry which of course is not applicable in

our case. The least effort principle also accounts for the fact that the most frequent words and those with the lowest rank tend to be short function words such as "the, and, of, but" etc. which are easy to coin and whose cost of usage is small.
Using the Zipf law of expression as a starting point, we derived word significance factors based on frequency characteristics of individual words in gene symbol's biological description text. We followed this approach to quantify a potential stop list: by observing more closely the term frequency we decided on which suitable high and low threshold value to remove all words with a collection frequency above and below this threshold. However, low and high-frequency terms might produce losses in precision and recall. Another problem is the necessity to choose appropriate thresholds in order to distinguish the useful medium-frequency term from the remainder. Therefore the rest of the list was manually checked to discard meaningful term from generic and non-function terms ***(see additional files, list stop words)***.

Gene symbol's sense discrimination
Data clustering (Vector space representation of gene symbols).
Figure 6 "step 2" highlights the sense discrimination stage. As the current text mining technologies are not able to "read" and "understand" a text like human beings due to its unstructured nature, text mining transformed texts into a vector space model, to which existing data mining or machine learning algorithms can easily be applied. In a vector space model, a text is represented as a vector by means of representative keywords called index terms.
A useful index term must fulfill a dual function: it must be related to the information content of the gene symbol, so as to render the item retrievable when it is wanted (recall); a good index term also distinguishes the gene symbols to which it is assigned from the remainder to prevent the indiscriminate retrieval of all items, whether wanted or not (precision). Thus, a term such as protein, gene, RNA, DNA is not very indicative of the potential biological function but rather of a certain class of biological molecules. This suggests the use of relative frequency measures to identify terms occurring with substantial frequencies in some individual gene symbols of a collection, but with a relatively low overall collection frequency. Such terms may then help in retrieving the items to which they are assigned, while also distinguishing them from the remainder in the collection.
The index terms describing a gene symbol were those from UniProtKB description field (DE), Cross-reference fields (DR) and comment fields (CC), devoid of those present in the Stop list generated previously. Therefore, for the purpose of further clustering a gene symbol's corpus (all document) was constructed. A document is made up of a species and gene symbol description index terms pair.
The aim of document clustering is defined as follows: Given a set of n documents called D_S, D_S is clustered into a user-defined number of k document cluster D_{S1}, D_{S2},... D_{Sk} (ie $\{D_{S1}, D_{S2}, \ldots D_{Sk}\} = D_S$) so that the documents in a document cluster are similar to one another while documents from different clusters are dissimilar. One innovation in our pipeline was to predict the optimal number of clusters given a dataset without user intervention as the unsupervised task of document clustering is a very subjective task. One important question that might arise from the latter is how can we predict the optimal number of cluster given a data set?
Gordon and colleagues (39) gave a good overview on many methods that have been proposed for estimating the optimum number of clusters given a dataset. These methods could be classified as global or local. The former evaluate some measures over the entire dataset and optimize it as a function of the number of clusters. The local method considers individual pairs of clusters and tests whether they should be merged.

Robert Tibshirani and colleagues (40) applied six different methods for estimating the number of clusters. Their stimulation studies suggested that the Gap statistics estimate is a good algorithm to identify well-separated clusters. Given that biological terms co-occurrence in the description of the biological function of a gene can be clearly separated, we believe this method as well suited for clustering and thus discriminating biological function of ambiguous gene symbols efficiently.

The Gap statistic relied on the following: If one tries to cluster a dataset (i.e. numerous observations described in terms of a feature space) into n groups or clusters and if we plot the graph of within cluster dissimilarity (error) or similarity along an Y-axis and the number of clusters along an X-axis, then this graph generally takes a form of an elbow or knee depending upon the measure on the Y-axis. The Gap statistic seeks to locate this elbow or knee because the x-value of this elbow represents the optimal number of clusters for the data set. In ***Figure 7 "Gap statistic curve of ALK1" (see additional files)*** the Gap curve of ***ALK1*** shows many local maxima and these in itself can be informative and suggesting the very subjective native of document clustering.

In our methodology we integrated and used a suite of clustering algorithm, namely CLUTO for clustering(41) .

Clustering approaches can be categorized as hierarchical, partitional (42) and hybrid. Next we will give a brief overview on clustering techniques and the choice of those used throughout our methodology.

Hierarchical agglomerative clustering algorithms successively merge the most similar objects based on pair-wise distances between the objects until a termination condition holds (criterion function). Criterion function is a term that refers to different metrics that clustering algorithm use to try to optimize the quality of a clustering solution. An advantage of hierarchical agglomerative algorithm is that they generate a document hierarchy that users can search up and down for specific topics of interest. However, due to their cubic time complexity, they are limited for a very large number of documents.

Partitional clustering algorithm, most widely used, first randomly select the k centroid and then divide the object into k disjoint groups through iteratively relocating objects based on the similarity between the centroids and the object. Hence, partitional techniques display a linear time complexity. One major drawback of partitional algorithm is that clustering results are heavily sensitive to the initial centroid because the centroids are randomly selected.

Hybrid clustering algorithm is a partitional method that produces hierarchical clustering solutions using repeated bisections. The intention is to take advantage of the global view of the partitional algorithm but also to reduce the instability induced by the initial random k centroid. An example of hybrid clustering algorithm is the repeated bisection.

Criterion functions are classified into internal, external and hybrid type. The internal type takes an intra-cluster (within) view of the clustering process, thus only captures how the gene symbol's context-vector in any given cluster is related to each other.

$$I1 = \sum_{r=1}^{k} n_r \left(\frac{1}{n_r^2} \sum_{d_i, d_j \in S_r} \cos(d_i, d_j)\right) \qquad (1)$$

Where $I1$ is the internal maximization function, n_r represents the size of each k cluster (gene symbol biological functions), $cos(d_i,d_j)$ is the similarity measure between context-vector d_i and d_j

The external criterion functions take an inter-cluster (between) view and try to find gene symbol's biological function clusters which are as different or dissimilar from each other as possible.

$$E1 = \sum_{r=1}^{k} n_r \cos(C_r, C) \qquad (2)$$

Where E1 is the external minimization criterion function, minimizing the similarity ($\cos(C_r,C)$) between the centroid of each cluster (C_r) and the centroid of the entire set (C) weighted by the size of each cluster (n_r)
The hybrid criterion function strives to propose clustering solutions which maximize intra-cluster similarity and simultaneously minimize the inter-cluster similarity.

$$H1 = \frac{I1(m)}{E1(m)} \qquad (3)$$

The Hybrid criterion function *H1* is proportional to a maximization function and inversely proportional to a minimization function and therefore itself is a maximization function.
Zhao and Karypsis (43) reported that a strength of repeated bisection over agglomerative algorithm is its aptitude for excluding merging errors in the early stage that are usually the cause of poor performance in the agglomerative method. Furthermore, Pundare and Pederson (44) conjectured that when sparse data are available (specifically in this study as the corpus for each symbol prior to analysis were very sparse), repeated bisection can improve the clustering process. Owing to the advantage hybrid criterion function and hybrid clustering procedure offers we used them in our entire methodology.

Cluster labeling
Once the context-vector was separated into clusters by CLUTO repeated bisection as clustering procedure and hybrid criterion function, CLUTO generated a set of descriptive and discriminating features based on a set threshold of most characteristic features that were unique to each cluster. These were the top N (10 in our case) grams (words) ranked on their frequency or their statistical scores. The idea here was to assign automatically the most significant words summarizing a cluster without having to examine the clusters content. We would like to emphasize that these summaries were simple word lists without any grammatical syntax.
Furthermore, to each cluster we assigned the latest common ancestor of species within that clusters using tools from the BioPerl toolkit (45). Taxonomic entries were from the NCBI taxonomic database, which is cross-linked from UniProtKB entries.

Data Storage and querying
All data generated throughout our pipeline is loaded and stored in a MySQL database for further querying. Sample table headers that show the link between gene symbols, cluster ID and latest common ancestor LCA is shown in ***table 2.*** This data are paramount in the subsequent step of our disambiguation because they are used to train a classifier for future classification of an ambiguous gene symbol.

Gene symbol	Kingdom	Tax rank	Function/ #species	# of cluster/ cluster size (species)
ALK1	Bacteria	Bacillus cereus	2;1	3;3
ALK1	Fungi	Ascomycota	1;4	3;4
ALK1	Fungi	Aspergillus	0;3	3;3
ALK1	Fungi	Saccharomyces cerevisiae	2;1	3;3
ALK1	Metazoa	Homo sapiens	2;1	3;3
CAT2	Archaea	Halobacteriaceae	1;2	2;17
CAT2	Bacteria	Bacteria	1;8	2;17
CAT2	Fungi	Ascomycota	1;5	2;17
CAT2	Fungi	Pezizomycotina	0;2	2;21
CAT2	Metazoa	Mayetiola_destructor	0;1	2;21
CAT2	Metazoa	Rattus_norvegicus	1;1	2;17
CAT2	Viridiplantae	Magnoliophyta	0;18	2;21
CAT2	Viridiplantae	Populus_deltoides	1;1	2;17

Table 2: Data set used for training the classifier in the case of ALK1 and CAT2. Columns 1-3 are self-descriptive. Column 4 contains 2 digits, the first represents the cluster identifier, the second the number of species in that cluster having the taxonomic rank in column 3. Column 5 summarizes the number of clusters for a specific symbol followed by the total cluster size.

Gene symbol's sense disambiguation

Figure 6 "step 3" highlights this stage

Word Sense Disambiguation (WSD) may be seen as a text categorization (TC) task (46), which is the task of assigning a Boolean value to each pair $(d_j, c_i) \in DxC$, where D is a domain of the document and $C = \{c_1, ..., c_{|c|}\}$ a set of predefined categories. In our case, once we have viewed the gene symbol occurrence context as a document and gene function as a category, this is a single-label TC, and one in which document-pivoted TC is usually the right choice. We should view TC in our approach as taxonomy-pivoted gene function categorization.

WSD are usually coupled to machine learning (ML) techniques. In ML terminology the classification problem is an activity of supervised learning, since the learning process is "supervised" by the knowledge of the categories and the training instances that belong to them. Amongst various existing machine learning techniques such as Support vector machine (SVM), we have chosen the probabilistic classifier Naïve Bayes classifier because of its simple and yet robust implementation. The Naïve Bayes classifier assumes that any two coordinates of the document vector are, when viewed as random variables, statistically independent of each other. The naïve character of the classifier is due to the fact that the latter assumption is quite obviously not verified in practice. This probabilistic classifier is mathematically defined as follow:

$$P(c_i \mid \vec{d}_j) = \frac{P(c_i)P(\vec{d}_j \mid c_i)}{P(\vec{d}_j)}$$

In the above formula the event space is that of a specific ambiguous gene symbol.

$P(\vec{d}_j)$ is the probability that a randomly selected LCA (latest common ancestor) encompasses species representing vector $\vec{d}_j$. And $P(c_i)$ the probability that a randomly picked species belongs to category c_i therefore has function c_i.

In order to evaluate the approach analytically and to prove that our system was correct and complete, we needed a formal specification of the problem the system was trying to solve. Due to its inherent subjective character it is difficult to formalize the notion of text categorization. Therefore our evaluation was conducted experimentally rather than analytically.

Our experimental evaluation measured the ability of the system to take the right classification decision (effectiveness) through precision and recall. The precision (p) is defined as the probability that, if a random gene symbol-species pair (d_x) is classified under biological function (c_i), this decision is correct. Mathematically as:

$$p = \frac{TP}{TP + FP} \qquad \text{(I)}$$

Where TP= number of true positive decision, FP= number of false positive decision.

The recall (r) is defined as the probability that, if a random gene symbols- species pair (d_x) ought to be classified under biological function (c_i), this decision is taken. Mathematically as:

$$r = \frac{TP}{TP + FN} \qquad \text{(II)}$$

Where TP= number of true positive decision, FN= number of false negative decision.

Therefore, precision may be regarded as the "degree of soundness" and recall as the "degree of completeness".

Acknowledgements

The authors wish to thank Gert Vriend, Monica Chagoyen and Lynette Hirschman for valuable critics and feedback and Jeannette Kluess for careful reading of the manuscript. This project was (co) financed by the Centre for BioSystems Genomics (CBSG) which is part of the Netherlands Genomics Initiative/Netherlands Organization for Scientific Research.

References

1. MacMullen, W.J. and Denn, S.O. (2005) Information Problems in Molecular Biology and Bioinformatics. *Journal of the American society for information science and technology*, **56**, 447-456.
2. Swanson, D.R. and Smalheiser, N.R. (1999) Implicit text linkages between Medline records: Using Arrowsmith as an aid to scientific discovery. *Library Trends*, **48**, 48-59.
3. Swanson, D.R. (1990) Somatomedin C and Arginine: Implicit connections between mutually isolated literatures. *Perspectives in Biology and Medicine*, **33**, 157-179.
4. Swanson, D.R. (1988) Migraine and magnesium: eleven neglected connections. *Perspectives in Biology and Medicine*, **31**, 526-557.
5. Swanson, D.R. (1987) Two medical literatures that are logically but not bibliographically connected. *JASIS*, **38**, 228-233.
6. Swanson, D.R. (1986) Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspectives in Biology and Medicine*, **30**, 7-18.

7. Yuryev, A., Mulyukov, Z., Kotelnikova, E., Maslov, S., Egorov, S., Nikitin, A., Daraselia, N. and Mazo, I. (2006) Automatic pathway building in biological association networks. *BMC bioinformatics*, **7**, 171.
8. Narayanasamy, V., Mukhopadhyay, S., Palakal, M. and Potter, D.A. (2004) TransMiner: mining transitive associations among biological objects from text. *J Biomed Sci*, **11**, 864-873.
9. McDonald, D.M., Chen, H., Su, H. and Marshall, B.B. (2004) Extracting gene pathway relations using a hybrid grammar: the Arizona Relation Parser. *Bioinformatics (Oxford, England)*, **20**, 3370-3378.
10. Donaldson, I., Martin, J., de Bruijn, B., Wolting, C., Lay, V., Tuekam, B., Zhang, S., Baskin, B., Bader, G.D., Michalickova, K. *et al.* (2003) PreBIND and Textomy--mining the biomedical literature for protein-protein interactions using a support vector machine. *BMC bioinformatics*, **4**, 11.
11. Alako, B.T., Veldhoven, A., van Baal, S., Jelier, R., Verhoeven, S., Rullmann, T., Polman, J. and Jenster, G. (2005) CoPub Mapper: mining MEDLINE based on search term co-publication. *BMC Bioinformatics*, **6**, 51.
12. Venclovas, C., Zemla, A., Fidelis, K. and Moult, J. (2003) Assessment of progress over the CASP experiments. *Proteins*, **53 Suppl 6**, 585-595.
13. Wodak, S.J. and Mendez, R. (2004) Prediction of protein-protein interactions: the CAPRI experiment, its evaluation and implications. *Curr Opin Struct Biol*, **14**, 242-249.
14. Johnson, K.F. and Lin, S.M. (2001) Critical assessment of microarray data analysis: the 2001 challenge. *Bioinformatics (Oxford, England)*, **17**, 857-858.
15. Reese, M.G., Hartzell, G., Harris, N.L., Ohler, U., Abril, J.F. and Lewis, S.E. (2000) Genome annotation assessment in Drosophila melanogaster. *Genome research*, **10**, 483-501.
16. Hersh, W.R., Bhupatiraju, R.T., Ross, L., Roberts, P., Cohen, A.M. and Kraemer, D.F. (2006) Enhancing access to the Bibliome: the TREC 2004 Genomics Track. *J Biomed Discov Collab*, **1**, 3.
17. Cohen, A.M. and Hersh, W.R. (2006) The TREC 2004 genomics track categorization task: classifying full text biomedical documents. *J Biomed Discov Collab*, **1**, 4.
18. Hirschman, L., Yeh, A., Blaschke, C. and Valencia, A. (2005) Overview of BioCreAtIvE: critical assessment of information extraction for biology. *BMC bioinformatics*, **6 Suppl 1**, S1.
19. Hirschman, L., Colosimo, M., Morgan, A. and Yeh, A. (2005) Overview of BioCreAtIvE task 1B: normalized gene lists. *BMC bioinformatics*, **6 Suppl 1**, S11.
20. Colosimo, M.E., Morgan, A.A., Yeh, A.S., Colombe, J.B. and Hirschman, L. (2005) Data preparation and interannotator agreement: BioCreAtIvE task 1B. *BMC bioinformatics*, **6 Suppl 1**, S12.
21. Blaschke, C., Leon, E.A., Krallinger, M. and Valencia, A. (2005) Evaluation of BioCreAtIvE assessment of task 2. *BMC bioinformatics*, **6 Suppl 1**, S16.
22. Yeh, A., Morgan, A., Colosimo, M. and Hirschman, L. (2005) BioCreAtIvE task 1A: gene mention finding evaluation. *BMC bioinformatics*, **6 Suppl 1**, S2.
23. Chen, L., Liu, H. and Friedman, C. (2005) Gene name ambiguity of eukaryotic nomenclatures. *Bioinformatics*, **21**, 248-256.
24. Schijvenaars, B.J., Mons, B., Weeber, M., Schuemie, M.J., van Mulligen, E.M., Wain, H.M. and Kors, J.A. (2005) Thesaurus-based disambiguation of gene symbols. *BMC Bioinformatics*, **6**, 149.

25. Xu, H., Markatou, M., Dimova, R., Liu, H. and Friedman, C. (2006) Machine learning and word sense disambiguation in the biomedical domain: design and evaluation issues. *BMC bioinformatics*, **7**, 334.
26. Gaudan, S., Kirsch, H. and Rebholz-Schuhmann, D. (2005) Resolving abbreviations to their senses in Medline. *Bioinformatics (Oxford, England)*, **21**, 3658-3664.
27. Ginter, F., Boberg, J., Jarvinen, J. and Salakoski, T. (2004) New techniques for disambiguation in natural language and their application to biological text. *Journal of machine learning Research*, **5**, 605-621.
28. Hatzivassiloglou, V., Duboue, P.A. and Rzhetsky, A. (2001) Disambiguating proteins, genes, and RNA in text: a machine learning approach. *Bioinformatics*, **17 Suppl 1**, S97-106.
29. Liu, H., Lussier, Y.A. and Friedman, C. (2001) Disambiguating ambiguous biomedical terms in biomedical narrative text: an unsupervised method. *J Biomed Inform*, **34**, 249-261.
30. Podowski, R.M., Cleary, J.G., Goncharoff, N.T., Amoutzias, G. and Hayes, W.S. (2004) AZuRE, a scalable system for automated term disambiguation of gene and protein names. *Proc IEEE Comput Syst Bioinform Conf*, 415-424.
31. Schuemie, M.J., Kors, J.A. and Mons, B. (2005) Word sense disambiguation in the biomedical domain: an overview. *J Comput Biol*, **12**, 554-565.
32. Alako, B.T., Rainey, D., Nijveen, H. and Leunissen, J.A. (2006) TreeDomViewer: a tool for the visualization of phylogeny and protein domain structure. *Nucleic Acids Res*, **34**, W104-109.
33. Tamames, J. and Valencia, A. (2006) The success (or not) of HUGO nomenclature. *Genome Biol*, **7**, 402.
34. Petsko, G.A. (2002) What's in a name? *Genome Biol*, **3**, COMMENT1005.
35. Pearson, H. (2001) Biology's name game. *Nature*, **411**, 631-632.
36. Kestler, H.A., Muller, A., Gress, T.M. and Buchholz, M. (2005) Generalized Venn diagrams: a new method of visualizing complex genetic set relations. *Bioinformatics (Oxford, England)*, **21**, 1592-1595.
37. Bairoch, A., Apweiler, R., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M. *et al.* (2005) The Universal Protein Resource (UniProt). *Nucleic acids research*, **33**, D154-159.
38. Zipf G, K. (1949) *Human behaviour and the principle of least effort.* Addision Wesley Publishing, Reading, Massachusetts.
39. Gordon, A. (1999) *Classification (2nd edition).* Chapman and Hall/CRC press, London.
40. Tibshirani, R., Walther, G. and Hastie, T. (2001) Estimating the number of clusters in a dataset via the Gap statistic. *Journal of the Royal Statistics Society (Series B)*, **63**, 411-423.
41. Karypis, G. (2003).
42. Kaufman, L. and Rousseeuw, P.J. (1999) *Finding groups in Data: an Introduction to Cluster Analysis.* John Wiley & Sons.
43. Zhao, Y. and Karypis, G. (2002) Evaluation of hierarchical clustering algorithms for document datasets. *In proceedings of the 11th conference of information and knowledge management (CIKM)*, 515-524.
44. purandare, A. and and Pedersen, T. (2004) Word sense discrimination by clustering context in vector and similarity space. *In Proceedings of the CoNLL*, 41-48.

45. Stajich, J.E., Block, D., Boulez, K., Brenner, S.E., Chervitz, S.A., Dagdigian, C., Fuellen, G., Gilbert, J.G., Korf, I., Lapp, H. *et al.* (2002) The Bioperl toolkit: Perl modules for the life sciences. *Genome research*, **12**, 1611-1618.
46. Escudero, G., Marquez, L. and Rigau, G. (2000), *Proceedings of ECML-00, 11th European Conference on Machine Learning* Barcelona, Spain, pp. 129-141.

Additional files

Additional file 1 – Correlation between phenotype (different senses per ambiguous gene symbol) and genotype (TreeDomViewer sequence data analysis)
Link to files supporting correlation between phenotype (Textual) and genotype (Sequences) per ambiguous gene symbols randomly selected. ("See table through the link", PDF, www.bioinformatics.nl/tools/gi/additional/).

Additional file 2 – Leave-one-out cross-validation of species gene-symbol pairs GI functional assignment with respective Naïve Bayes probabilities Phenotype (index textual data) of different senses addressing each randomly selected gene symbol ("leave_one_out_crossvalidation.txt", TXT, www.bioinformatics.nl/tools/gi/additional/).

Additional file 3 – SwissProt descriptions of the 100 randomly selected gene symbols used in the performance evaluation. ("gene_symbols_SwissProt_description.txt", TXT, www.bioinformatics.nl/tools/gi/additional/)

Additional file 4 - Performance metric on 100 randomly selected gene symbols. ("Performance_metric.txt", TXT, www.bioinformatics.nl/tools/gi/additional/)

Additional file 5 – Gap statistic curve of ALK1 ("Gap_statistics_curve_of_alk1.png", PNG, www.bioinformatics.nl/tools/gi/additional/)

Additional file 6 – List of stop words, this mainly common word, species names, numeric character, chemical compound and non-alphanumeric characters. ("Stop_list.txt", TXT, www.bioinformatics.nl/tools/gi/additional/)

Chapter 5

GeneIlluminator: disambiguation of PubMed abstracts

Blaise T.F Alako[1,2], Pieter B.T. Neerincx[1], Jack A.M. Leunissen[1]

[1]Laboratory of Bioinformatics, Wageningen University and Research Centre, PO Box 8128, 6700 ET Wageningen, The Netherlands,

[2]Centre for BioSystems Genomics, PO Box 98, 6700 AB Wageningen, The Netherlands

In preparation for submission

Abstract

An important consequence of the accumulation of vast amounts of genomic data is that the ambiguity of gene nomenclature leads to confusion in annotation. We present here GeneIlluminator, a tool that can highlight the multiple biological functions assigned to a gene within or across species exhibiting an ambiguous name or gene symbol. Additionally our tool can annotate a gene symbol of a given species based on prior knowledge of its closest taxonomic relative. This is an asset for annotation pipelines as well as document-pivoted and category-pivoted text categorization where gene symbols or gene name abbreviations are ambiguous. GeneIlluminator also proposes unambiguous gene symbol synonyms to the initial abbreviation of interest for a biological function. The suggested sets of unambiguous synonyms and biological entities of the category are used in a Boolean or vector model to effectively retrieve PubMed abstracts through GoPubMed, thus actively disambiguating PubMed abstracts. GeneIlluminator is freely available for academic use at: www.bioinformatics.nl/tools/gi/. For automated querying via custom software, four BioMOBY web services are available for remote programmatic access at: https://www.bioinformatics.nl/phenolink/home/BIF_services/GeneIlluminator_services.html.

INTRODUCTION

A consequence of the recent application of computational techniques in the life sciences is a vast increase in available data and publications. Naturally this provides a new and important source of valuable information, which however is presented in a challenging format, i.e. natural language text. In order to meet this challenge in large-scale exploitation, new text mining techniques are required. Several tools have been already developed to help researchers to extract and mine data in scientific literature, such as (1-6). However, barriers to successful selection and identification of gene names and biological terms are: the extensive lexical variation preventing terms to be recognized in a free text, the gene name synonymy and the gene name homonymy. The latter creates uncertainties regarding the exact identity of a term. Furthermore, the biological field is mined with a constantly changing terminology and constant additions to this terminology. A related problem is the lack of a stringent nomenclature in the majority of gene and protein databases. For example, the guidelines for FlyBase, dealing with the genome of Drosophila, is largely unrestricted (7). FlyBase favours a rather descriptive nomenclature, which makes an automated identification of gene names very difficult. In contrast, term conventions for yeasts are more stringent, thus allowing for easier gene name identification.

To this problem of rather technical ambiguity adds also the often encountered conflict of interest between researchers, as scientists might rather share their toothbrush than the same gene name (8).

Applications such as manual literature search, automated text-mining, named entity recognition, gene or protein annotation, and linking of knowledge from different information sources require the knowledge of all names referring to a given gene or protein unambiguously (9). In this context it is important to realize that biomedical and biological terms often appear in abbreviated forms, so-called acronyms. Although several methods have been developed to capture the different acronyms in the literature (10-19), they are not sufficient in selecting the proper acronym that unambiguously pertains to the concept of interest, e.g. the biological function. That means that to date current text mining tools cannot guide users toward the effective term selection to achieve a meaningful query.

This paper addresses the ambiguity at the gene nomenclature abbreviation or acronym level. How does one efficiently and unambiguously tap the huge knowledge base of literature in such a way as to assist scientist to make sense of the vast amount of high-throughput data generated in experiments? Good interpretation is key to generate new hypotheses for further experimentation and validation.

Here we present GeneIlluminator (GI), a tool that addresses various aspects of text mining. It addresses and displays the multiple aspects of the biological functions of ambiguous gene symbols. These multiple aspects can be used independently to partition PubMed abstracts based on a similarity profile between abstracts and different concepts of an ambiguous gene symbols. The tool can also be used as a gene symbol ambiguity checker for information contained in UniProtKB. It categorizes ambiguous gene symbols, their synonyms and species with respect to the distinct biological concept of the primary symbols used for the searches. Furthermore there is an option to check the quality of the categorization as well as the provision of the feature space used for the categorization task. And finally, GeneIlluminator can be seen as an interactive curation tool, which supports the curator of databases and can eventually learn from him. Besides this it represents a functional annotation approach that combines data from linguistic and bioinformatics sources. In the subsequent section we will introduce the methodology of GeneIlluminator.

MATERIALS AND METHODS

Implementation and design

Data preparation and processing

GeneIlluminator interrogates a MySQL repository of ambiguous gene symbols taken from a quantitative survey of ambiguous acronyms in UniProtKB. Our database was constructed with a methodology (manuscript submitted for publication) based on the UniProt Knowledgebase, release 12.2. Briefly, the UniProt consortium is concerned with the integration of protein information in the UniProtKB, providing a central, stable, comprehensive, richly classified and accurately annotated protein sequence database with extensive cross-references to other data sources. The expectation of the UniProt project is that the SwissProt/UniProt and Entrez gene databases will increasingly share nomenclature with the advantage that the mapping between databases will be increasingly complete and unambiguous. This will aid in facilitating the generation of gene name dictionaries which in turn will represent a comprehensive source of gene nomenclature for analysis and text mining purposes. (20)

Input and output description

The GI interface uses as the minimum input the gene name symbol or acronym. The interface uses AJAX) technologies that help to auto-complete input data such as gene symbol or species name. ***Figure 1*** displays the GI interface.

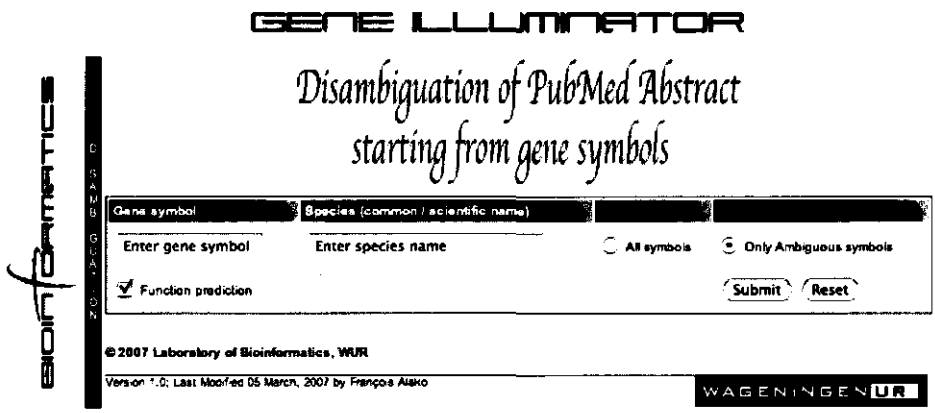
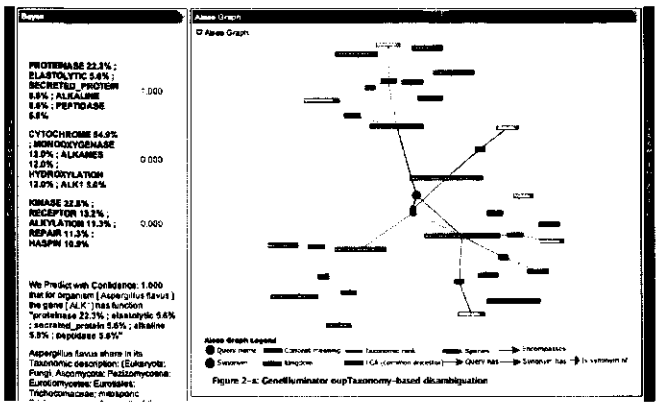


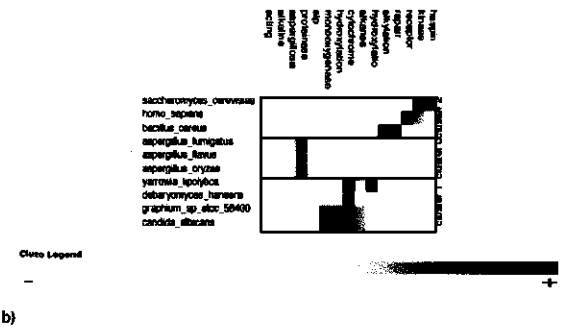
Figure 1: GeneIlluminator web interface

It uses aiSee, a graphic layout software tool (www.aisee.com) and CLUTO (http://glaros.dtc.umn.edu/gkhome/views/cluto) to enhance the visualization of GeneIlluminator's gene symbol analysis. GI exploits aiSee's force-directed layout to display the gene symbol and its synonyms in a network manner to the most descriptive terms of its biological function and different taxonomic level of the respective species. The nodes of the generated graph represent the gene symbol, its synonyms and associated biological terms, and the taxonomy.

Different shapes and colors are applied to provide a comprehensive overview to analyze this graph. All nodes are interactive and by clicking on the node a set of related PubMed abstracts will be displayed for user's assessment. The graph is illustrated as a publication quality scalable vector graphic (SVG) in ***Figure 2-a***, where the network of gene symbols and their synonym associations to descriptive biological terms are shown. Furthermore, the clustering toolkit CLUTO (***Figure 2-b***) provides clustering of descriptive features of the ambiguous symbol under investigation, which represents a complementary aspect to the graph layout of aiSee.



a)



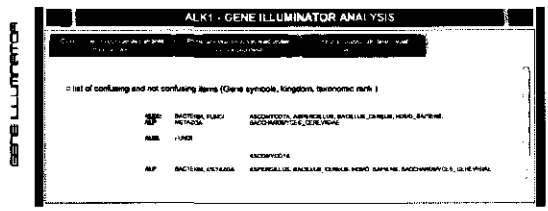
b)

Figure 2a-b, Geneilluminator output: (a) Taxonomy-based disambiguation of ALK1. (b) phenotypic (textual) data clustering of ALK1 annotation across different species showing three distinct biological functions.

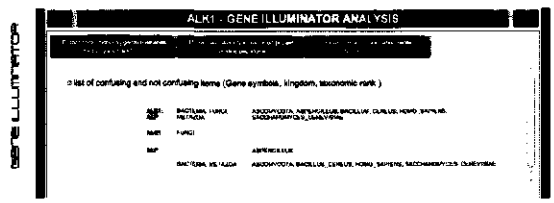
In the case that a gene symbol cannot be found for any user-specified species, GI will assign the meaning of the gene symbol to the species based on their latest common ancestor in the taxonomy. The latter assignment is realized with a Naïve Bayes classification algorithm that bases it assumption on the following: most assignments of annotation for molecular function rely, at least partially, on the assumption that genes with similar sequences also display similar biological functions.

This implies, that sequences are evolutionary related to a certain extent, the relationship denoting a common latest ancestor prior to the speciation event. Thus, species of a certain taxonomic rank most probably have the same biological function for a specific gene. Hence annotators copy information from orthologous sequences or more closely related species. We use this assumption in our approach to relate the gene symbol to a before unknown species.

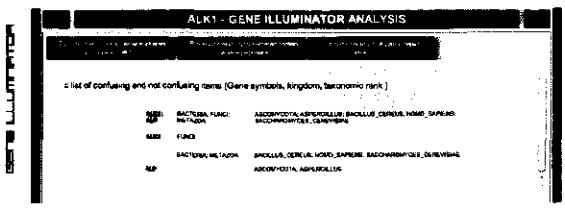
To illustrate our approach, we selected the gene symbol ALK1, which has entries in UniProtKB for human, fungi and bacteria domains. ***Figure 2-b*** shows the existence of two distinct groups in fungi, described respectively as cytochrome P450 and permease-peptidase related, whereas in bacteria ALK1 refers to DNA repair enzyme. Furthermore a comparison of ALK1 synonyms and taxonomic association to its multiple senses (***figure 2-c, 2-d, 2-e***) shows that unlike synonyms, the taxonomic units are unambiguous per biological function. For instance, species of the rank *Ascomycota* show unambiguously the function cytochrome monoxygenase related. All species of the taxonomic rank *Aspergillus* unambiguously have the function protease peptidase related, and last but not least *Bacillus cereus*, *Homo sapiens* and *Saccharomyces cerevisiae* unambiguously have the function kinase receptor related. This clearly exemplifies the issue of intra-species and inter-specie gene nomenclature ambiguity encountered in text mining of current knowledge databases.



c)



d)



e)

Figure 2c-d-e: GeneIlluminator output; (2-c), species distribution for the ALK1 biological function: Cytochrome monooxygenase related, (2-d) species distribution for the ALK1 biological function: proteinase peptidase related (2-e) species distribution for the ALK1 biological function: kinase receptor related.

Using the Naïve Bayes taxonomy-based disambiguation (***Figure 2-a***) shows that for the species *Aspergillus flavus* and the gene symbol ALK1 the function "proteinase 22.3%, elastolytic 5.6%, peptidase 5.6%" is assigned with a 100 % confidence based on the fact that all *Aspergillus spp.* already exhibited that function (latest common ancestor). The percentage ascribed to each functional biological term of a category, e.g. 22.3% proteinase, explains that the cluster is made up of "proteinase, elastolytic, peptidase, secreted protein, alkaline" to the extent of the given percentage. Currently predefined queries are generated from a list of synonyms as shown in ***figure 2-a, 2-c, 2-d, 2-e.*** A click on the cluster label 'Cytochrome monooxygenase cluster' retrieved a total of 157 abstract listed in PubMed with the majority of them referring or being related to fungi.

Design overview

GeneIlluminator is implemented in Perl as a web-based service, running on an apache 2.0 webserver using a Linux platform (SuSE Linux Enterprise Server 9 with MySQL 5.0). GI's interface (***Figure 1***) is a wrapper on several independent applications that uses a Naïve bayes algorithm for categorizing previously unseen instances of ambiguous gene symbols of a given species and subsequently plotting a graph of gene symbols, their synonyms and associated biological functions. The web interface preserves platform independency across multiple operating systems and allows the user to interact with the different GI programs without prior knowledge of computer programming skills. ***Figure 3*** summarizes a global overview of the GI workflow.
The GI web interface was tested on Windows XP, Mac OS X and several types of Linux OS browsers with good results. However, some problems where noticed with the interactive usage of the scalar vector graphics (SVG) due less or no support of some browsers with this graphic display; currently some browsers still require an Adobe SVG plug-in, downloadable from the Adobe site (http://www.adobe.com/svg/viewer/install/main.html). The latest versions of the Mozilla Firefox browser (version 2.0 and above) and Safari have already a native (built-in) SVG support and it is reasonable to expect that more browsers will soon follow.

In addition to the web interface, four BioMOBY (21) web services were developed, providing remote programmatic access to GI:

- GeneIlluminator_GetGraph
- GeneIlluminator_GetClusters
- GeneIlluminator_AssignSpeciesToCluster
- GeneIlluminator_GetPubMedQuery

These web services allow users to incorporate GI in workflows for automated disambiguation of gene symbols.

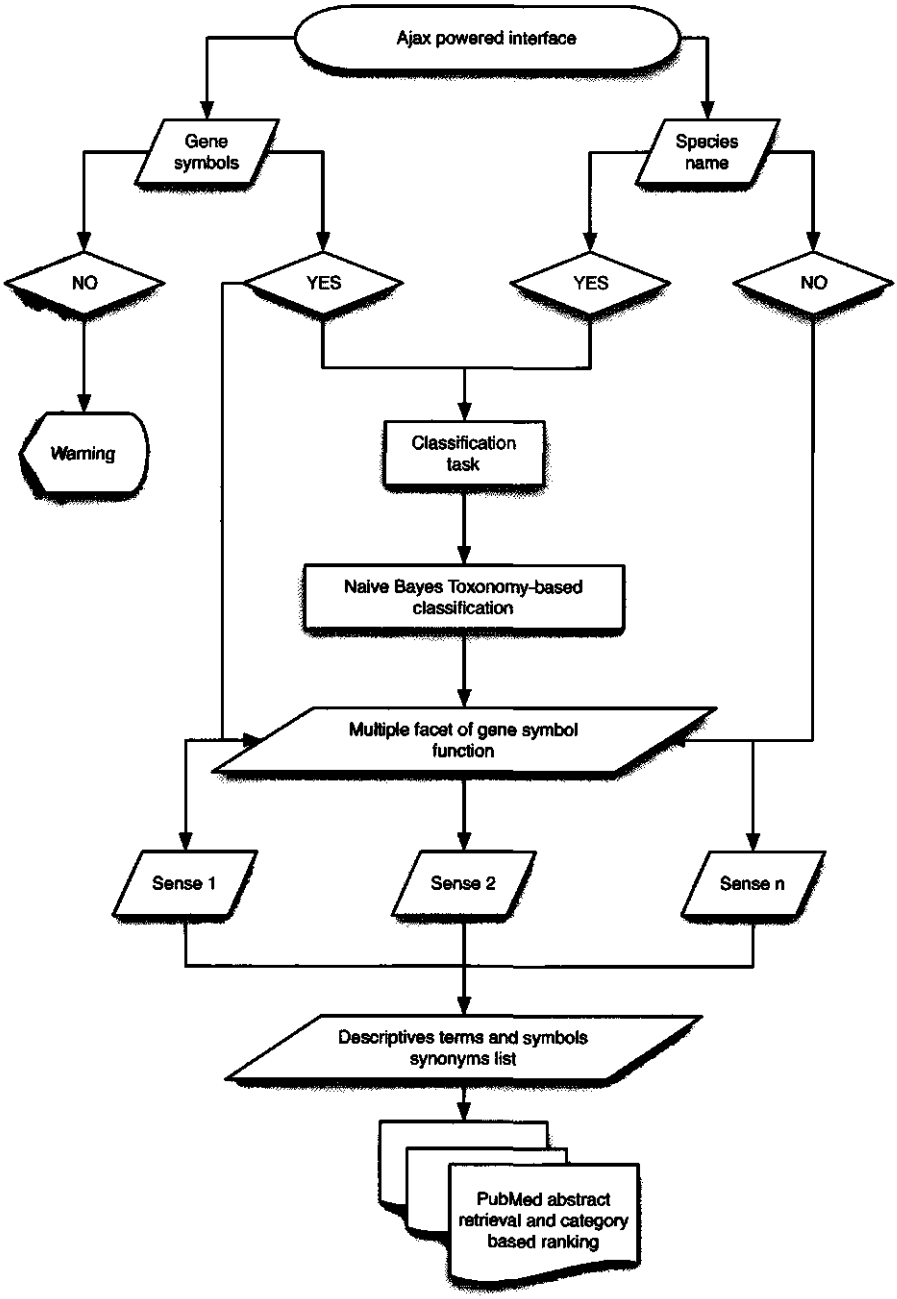


Figure 3: GeneIlluminator flow chart.

All services require a gene symbol as minimum input, but GeneIlluminator_AssignSpeciesToCluster and GeneIlluminator_GetPubMedQuery (***figure 4*** illustrate its Taverna workflow) require a species name in addition. The first two services provide an overview of all synonyms and homonyms for a gene symbol together with clusters of taxonomic clades showing the gene symbol pertaining to a certain function.

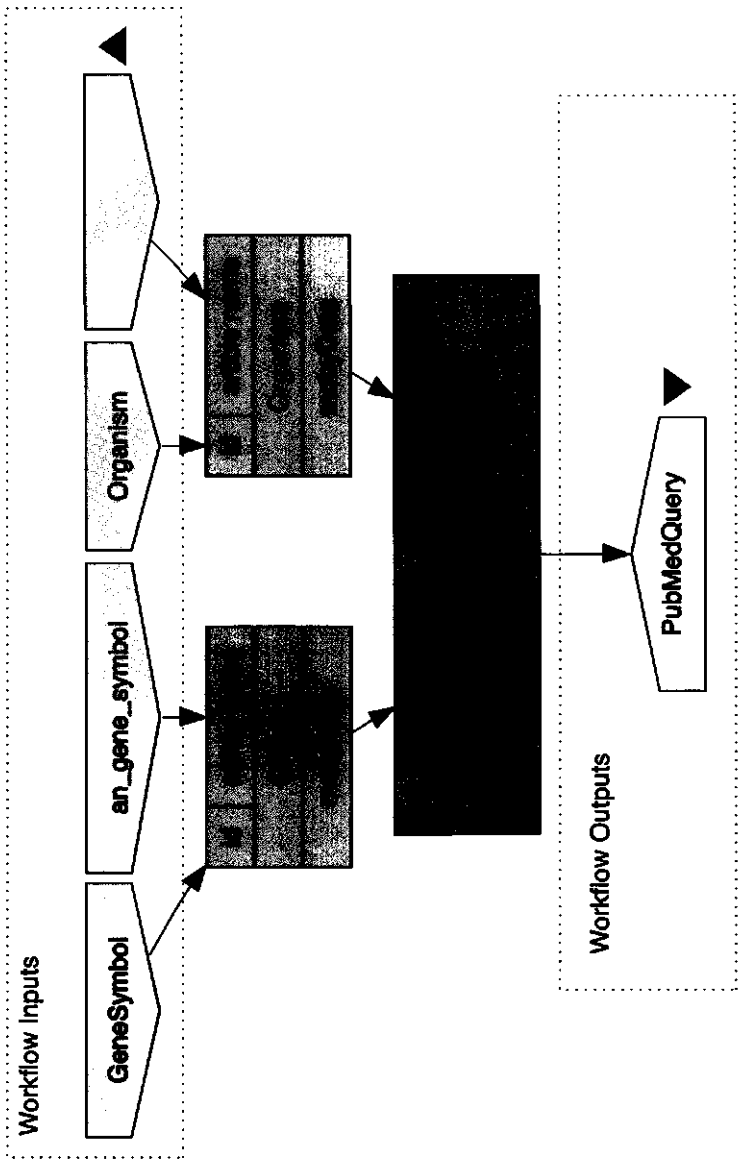


Figure 4: GeneIlluminator_GetPubMedQuery For an ambiguous gene symbol, this service use Gene Illuminator to create clusters describing which different genes sharing the same symbol exist in different parts of the tree of life. It then assigns the gene input symbol for the given input organism to one of the clusters Finally, using the cluster characteristics it creates a boolean PubMed query that could be used to unambiguously retrieve documents related to gene from the cluster the input gene symbol was assigned to.

GeneIlluminator_GetGraph provides this overview as an image in SVG format, whereas GeneIlluminator_GetClusters (***figure 5*** illustrates its Taverna workflow) provides the same information in a textual format (raw BioMOBY XML). GeneIlluminator_AssignSpeciesToCluster also provides the textual information as GeneIlluminator_GetClusters, but furthermore adds a Naïve Bayes probability indicating the likelihood of a gene symbol belonging to a certain cluster given the input species. Finally, GeneIlluminator_GetPubMedQuery implements a Boolean query to search PubMed for a given gene symbol and species, unambiguously retrieving documents that describe the gene of interest. It is noteworthy that GeneIlluminator_GetPubMedQuery provides the query terms solely and does not perform the actual query. The latter is accomplished by GoPubMed (22), a different web service software.
Documentation, example workflows and example inputs for the workflow builder Taverna (23) are available in the online material.

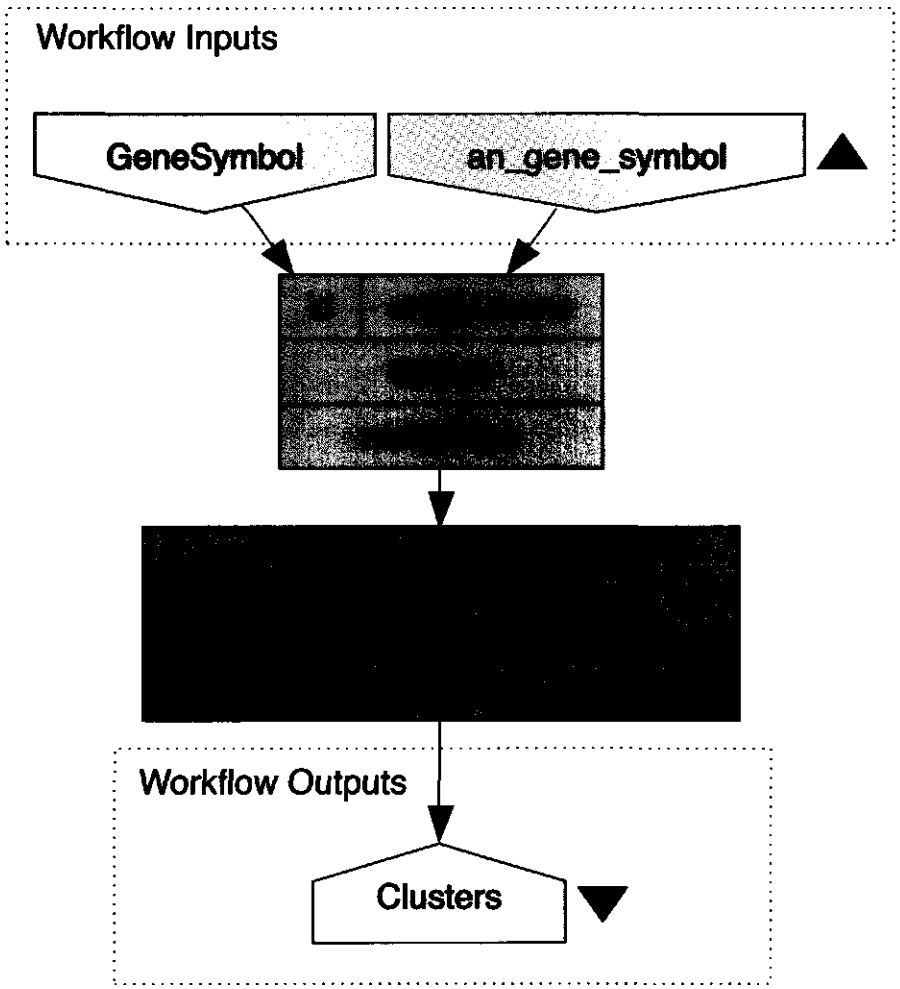


Figure 5: GeneIlluminator_GetClusters web service. If a gene symbol is ambiguous this service provides GI_Clusters describing which different genes sharing the same symbol exist in different parts of the tree of life Provides also gene symbol aliases associated to the input gene symbol

FUTURE PLAN

We plan in the near future to allow the user to select terms of categories and symbols to be used in her/his query formulation. This will gear up the search to the specific user needs. Furthermore we plan to provide a suitable interactive summary report of the different abstracts retrieved as well as association maps in suitable graphical format of all abstracts pertaining to each of the functional aspects of a gene symbol. Moreover we believe it to be beneficial to implement an ensemble of classifiers or use consensus from multiple classifiers to assign a function to an ambiguous gene symbol for a given species.

CONCLUSION

GeneIlluminator is a disambiguating text-mining tool that is able to display the multiple aspects, i.e. biological functions, of an ambiguous gene symbol. GI uses the latest common ancestor of a species assuming the same biological function for an ambiguous gene symbol, to infer the function in those species where the function is irretrievable or ambiguously retrievable through direct database query. Given a document, GI searches and retrieves all categories under which it should be filed (known as a document-pivoted categorization). Alternatively, given a specific category, GI searches and retrieves all the documents that should be filed under this specific category (category-pivoted categorization).
GeneIlluminator can be easily accessed through its web interface or its programmatic interfaces (web services) and represents a user-friendly tool for up-to-date text mining in life sciences.

Acknowledgment
The authors wish to thank Jeannette Kluess for careful reading of the manuscript, and Harm Nijveen for his help with programming the AJAX functionality of the user interface.
This project was financed by the Centre for BioSystems Genomics (CBSG) which is part of the Netherlands Genomics Initiative/Netherlands Organization for Scientific Research.

REFERENCES

1. Alako, B.T., Veldhoven, A., van Baal, S., Jelier, R., Verhoeven, S., Rullmann, T., Polman, J. and Jenster, G. (2005) CoPub Mapper: mining MEDLINE based on search term co-publication. *BMC Bioinformatics*, **6**, 51.
2. Liu, F., Jenssen, T.K., Nygaard, V., Sack, J. and Hovig, E. (2004) FigSearch: a figure legend indexing and classification system. *Bioinformatics*, **20**, 2880-2882.
3. Smalheiser, N.R. and Swanson, D.R. (1998) Using ARROWSMITH: a computer-assisted approach to formulating and assessing scientific hypotheses. *Comput Methods Programs Biomed*, **57**, 149-153.
4. Tanabe, L., Scherf, U., Smith, L.H., Lee, J.K., Hunter, L. and Weinstein, J.N. (1999) MedMiner: an Internet text-mining tool for biomedical information, with application to gene expression profiling. *Biotechniques*, **27**, 1210-1214, 1216-1217.
5. Tu, Q., Tang, H. and Ding, D. (2004) MedBlast: searching articles related to a biological sequence. *Bioinformatics*, **20**, 75-77.
6. Zhou, G., Wen, X., Liu, H., Schlicht, M.J., Hessner, M.J., Tonellato, P.J. and Datta, M.W. (2004) B.E.A.R. GeneInfo: a tool for identifying gene-related

biomedical publications through user modifiable queries. *BMC Bioinformatics*, **5**, 46.

7. Tuason, O., Chen, L., Liu, H., Blake, J.A. and Friedman, C. (2004) Biological nomenclatures: a source of lexical knowledge and ambiguity. *Pac Symp Biocomput*, 238-249.
8. Pearson, H. (2001) Biology's name game. *Nature*, **411**, 631-632.
9. Fundel, K. and Zimmer, R. (2006) Gene and protein nomenclature in public databases. *BMC Bioinformatics*, **7**, 372.
10. Adar, E. (2004) SaRAD: a Simple and Robust Abbreviation Dictionary. *Bioinformatics*, **20**, 527-533.
11. Chang, J.T., Schutze, H. and Altman, R.B. (2002) Creating an online dictionary of abbreviations from MEDLINE. *J Am Med Inform Assoc*, **9**, 612-620.
12. Frantzi, K., S. Ananiadou and H. Mima. (2000) Automatic recognition of multi-word Terms: The C-value/NC-value method. *International Journal on Digital Libraries*, **3**, 115-130.
13. Hisamitsu, T. and J. Tsujii. (2003), *Measuring Term Representiveness. Information Extraction in the Web Era*,. ed. M.T. Pazienza. ed. LNAI 2700, Springer, pp. 45-76.
14. Rimer, M. and O'Connell, M. (1998) BioABACUS: a database of abbreviations and acronyms in biotechnology and computer science. *Bioinformatics*, **14**, 888-889.
15. Rindflesch, T.C., Tanabe, L., Weinstein, J.N. and Hunter, L. (2000) EDGAR: extraction of drugs, genes and relations from the biomedical literature. *Pac Symp Biocomput*, 517-528.
16. Schwartz, A.S. and Hearst, M.A. (2003) A simple algorithm for identifying abbreviation definitions in biomedical text. *Pac Symp Biocomput*, 451-462.
17. Yoshida, M., Fukuda, K. and Takagi, T. (2000) PNAD-CSS: a workbench for constructing a protein name abbreviation dictionary. *Bioinformatics*, **16**, 169-175.
18. Yu, H., Hatzivassiloglou, V., Rzhetsky, A. and Wilbur, W.J. (2002) Automatically identifying gene/protein terms in MEDLINE abstracts. *J Biomed Inform*, **35**, 322-330.
19. Yu, H., Hripcsak, G. and Friedman, C. (2002) Mapping abbreviations to full forms in biomedical articles. *J Am Med Inform Assoc*, **9**, 262-272.
20. Apweiler, R., Bairoch, A., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M. *et al.* (2004) UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res*, **32**, D115-119.
21. Wilkinson, M.D. and Links, M. (2002) BioMOBY: an open source biological web services proposal. *Brief Bioinform*, **3**, 331-341.
22. Doms, A. and Schroeder, M. (2005) GoPubMed: exploring PubMed with the Gene Ontology. *Nucleic Acids Res*, **33**, W783-786.
23. Oinn, T., Addis, M., Ferris, J., Marvin, D., Senger, M., Greenwood, M., Carver, T., Glover, K., Pocock, M.R., Wipat, A. *et al.* (2004) Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics*, **20**, 3045-3054.

Chapter 6

Summary

Summary:

In the past years an extensive generation of information has taken place, in particular with the advent of the "omics", i.e. genomics, proteomics, transcriptomics, metabolomics and other members of the "omics" field, who use molecular techniques.
The "omics" field in current life sciences is characterized by the transition from local to global scale studies, as well as the generation of complete genomic sequences of unprecedented number and size which are currently stored and classified in specific data bases, e.g. UniProtKB. Another consequence of the rise of available data is their processing in scientific literature, which means that the number of publications increased exponentially alongside the development of the field.
In view of such extensive generation of information new automated approaches are required to aid in the organization and exploitation of these data sets in order to live up to the expectations of the "omics" rise. This requires advanced computational tools, mimicking some aspects of the manual interpretation process and thus computer science becomes an indispensable asset in assisting the automation of data analysis in biology. Therefore we have developed an integrative algorithmic approach towards unambiguous knowledge discovery in bioinformatics that is presented in this thesis.

We first introduce the field of Bioinformatics and computational biology in ***Chapter 1***. The step from local experimental approaches to global approaches successfully generates an overwhelming set of data and experimental results, thus advancing biology from technology and technique based research to information driven research. Consequently, computer techniques applicable to the biological fields are required to automate the organisation, management, and normalisation of such data sets and thus to assist scientists with the interpretation of results. Several biological fields where machine learning techniques find their application are introduced and a brief background to the respective evaluation techniques in machine learning are provided.
In general this chapter provides an overview on the main questions of bioinformatics and computational biology: definition of the field, importance of application, technical background of data generation, storage and organization as well as text mining and an outlook to the development of approaches in the subsequent chapters.

In ***Chapter 2*** we show that the traditional methods of inferring and supporting homology based on sequence similarity and identity might fail in the case of low sequence similarity. In order to solve the problem of low sequence similarity we developed TreeDomViewer (TDV), a biological web-based visualization tool that combines sequence alignment and InterProScan analysis of sequences and generates a phylogenetic tree projecting the predicted protein domains onto the multiple sequence alignment.
To illustrate the power of TreeDomViewer we used the lipocalins, a superfamily of proteins that carry hydrophobic prosthetic groups. Lipocalins have a strong divergent protein sequence, denoting a fast rate of molecular evolution. Moreover, the evolutionary history of the lipocalins is rich in gene duplication events, which increases the difficulty of obtaining an understanding of orthologous relationships. The results obtained with TreeDomViewer clearly show the relationships between the lipocalin subfamilies, where the alignment of the distinct domains underpin the phylogeny, and vice versa. It illustrates that TreeDomViewer helps in any phylogenetic analysis resolving both the

relationship among different group members and the relationship between groups, based solely on the aligned domain structure of each participant..

Reliable and robust interpretation of experimental data has been shown when those experimental data were integrated with other data source, namely the knowledge embedded in the literature. Therefore data integration is the key motto in this genomic era. ***Chapter 3*** uses this approach in the analysis of microarray data by introducing the tool CoPub Mapper for mining the literature based on term (gene names, disease, drug, chemical compound) co-occurrence. CoPub Mapper was developed using literature integration for the purpose of interpreting a set of differentially expressed genes generated from microarrays. These gene expression sets were selected from a comparison between ovaries of healthy women and women suffering from Polycystic Ovary Syndrome (PCOS) in an investigation of the causes of female infertility. CoPub mapper allows for a quick and versatile querying of co-published genes and keywords and was successfully used to cluster predefined groups of genes with their respective biological process, disease keyword and microarray data.

However, currently a directed literature information search or literature mining faces the problem of ambiguity. Ambiguity is inherent in natural language and its importance is evident given the fact that each publication uses natural language as the main vehicle of information distribution. Therefore we focussed on the ambiguity encountered in gene nomenclature, in particular on their use of abbreviation or acronyms of gene names (gene symbols).
The objective of ***Chapter 4*** is firstly to quantify the ambiguity problem in the universal protein knowledge base (UniProtKB) and secondly to propose a disambiguation approach based on species taxonomy. Given that sequences are evolutionary related to a certain extent, their relationship denoting a common latest ancestor prior to the speciation event, we hypothesize that gene symbols of taxonomically closely related species are more likely to be pointing to the same biological function for a specific gene. The latter hypothesis bases its foundation on the fact that biologists and annotators often copy names and functions from related species. Currently most assignments of a molecular function during annotation rely, at least partially, on this assumption. Using this key assumption we developed an algorithm that unambiguously assigns biological function to a given gene symbol and its alternative names, based on the latest common ancestor (LCA) of the given species name. The tool efficiently and unambiguously enriches query terms for searching the literature, starting solely with a gene symbol.

Chapter 5 introduces GeneIlluminator (GI), an application that implements the disambiguation methodology introduced in ***Chapter 4***. The calculated sets of unambiguous synonyms and their biological entities are implemented in a Boolean or vector model to effectively retrieve abstracts from the PubMed database with the aid of GoPubMed, an ontology-based PubMed search engine. Given a document, GI searches all categories under which the document in question could be filed; this method is known as document-pivoted categorization. Alternatively, given a specific category, GI searches all documents that should be filed under this category (category-pivoted categorization). Therefore, GeneIlluminator can be used to effectively disambiguate abstracts in the Medline database.

In conclusion

In the scope of this thesis we aimed to develop different approaches aiding in the utilization of the exponentially growing amount of information available in current life sciences. The main contributions and developments of this work thesis are:

1. TreeDomViewer, a visualization tool for phylogeny and protein domain structure, available at http://www.bioinformatics.nl/tools/treedom/
2. CoPub Mapper, a text mining tool based on term co-occurrence, available at http://services.nbic.nl/cgi-bin/copub/CoPub.pl
3. A taxonomy-based gene symbol disambiguation algorithm finding its application in literature retrieval for gene function prediction and in efficient document categorisation
4. GeneIlluminator, a tool for information retrieval and disambiguation of PubMed abstracts, available at: http://www.bioinformatics.nl/tools/gi/

Samenvatting

In de afgelopen jaren is er een massieve hoeveelheid informatie geproduceerd, met name met de opkomst van het "omics" onderzoeksveld, zoals genomics, proteomics, transcriptomics en metabolomics.
Het "omics" veld wordt gekarakteriseerd door een schaalvergroting in zowel breedte als diepte waardoor tot nu toe ongekende hoeveelheden data in databanken opgeslagen worden. Dit heeft weer tot gevolg dat het aantal wetenschappelijke publicaties exponentieel toeneemt met de ontwikkeling van het veld.
Om de organisatie en exploitatie van deze data het hoofd te kunnen bieden zijn nieuwe geautomatiseerde procedures onontbeerlijk. In dit proefschrift worden een aantal integratieve algoritmes gepresenteerd die behulpzaam zijn bij de eenduidige kennisvergaring in de bioinformatica.

In hoofdstuk 1 wordt het veld van de bioinformatica en computationele biologie geïntroduceerd. De schaalvergroting in de experimentele biologie leidt tot overweldigende hoeveelheden data en resultaten. Bijgevolg zijn informatica technieken noodzakelijk om het beheer, ontsluiten en analyseren van deze data mogelijk te maken, en de wetenschapper te ondersteunen in de interpretatie van de resultaten.
Dit hoofdstuk poogt een overzicht te geven van een aantal belangrijke vraagstukken in de bioinformatica, met name op het gebied van de "machine learning" en text mining, en de daarbij behorende technieken en evaluatie methodes.

In hoofdstuk 2 laten we zien dat de traditionele methodes voor het afleiden en ondersteunen van homologie gebaseerd op sequentie overeenkomst mogelijk faalt wanneer de sequentie overeenkomst erg laag is. Om dit probleem op te lossen ontwikkelden we TreeDomViewer (TDV), een web-gebaseerd visualisatie hulpmiddel dat sequentie alignment combineert met InterProScan eiwit domein analyse en fylogenetische analyse, waarbij de voorspelde domeinen op de multiple alignment geprojecteerd worden.
Om de kracht van TDV te illustreren werden de lipocalins gebruikt, een eiwit superfamilie die kleine hydrofobe moleculen zoals retinol kunnen binden. Lipocalins hebben een sterk gedivergeerde eiwit volgorde, wat wijst op een snelle moleculaire evolutie. Bovendien hebben er veel gen duplicaties plaatsgevonden, waardoor het berkrijgen van een begrip van de ortologe relaties bemoeilijkt wordt. De resultaten die met TDV verkregen worden tonen duidelijk de relaties tussen de verschillende lipocaline subfamilies, waarbij de alignment van de verschillende domeinen de fylogenie ondersteunt en *vise versa*. Het laat zien dat TDV behulpzaam is bij fylogenetische analyses voor het oplossen van de relaties tussen verschillende groepsleden als wel tussen groepen onderling, daarbij slechts gebruik makend van de ge-aligneerde domeinstructuur van elk van de eiwitten.

Data integratie is de benadering die in hoofdstuk 3 gebruikt wordt voor de analyse van microarray data. CoPub Mapper is een programma om de literatuur te doorzoeken gebaseerd op het gekoppeld voorkomen van termen (gen namen, ziektes, geneesmiddelen, chemische stoffen). Het programma is ontwikkeld met als doel het interpreteren van sets van differentieel tot expressie gebrachte genen van microarrays. Deze sets waren afkomstig een vergelijking tussen ovaria van gezonde vrouwen en vrouwen die aan Polycystic Ovary Syndrome (POS) leden, in een onderzoek naar de oorzaken van onvruchtbaarheid. CoPub Mapper maakt het mogelijk om snel en flexibel te zoeken naar ge-co-publiceerde genen en trefwoorden, en was succesvol in het

clusteren van groepen genen met hun bijbehorende biologische processen, ziekte termen en microarray gegevens.

Een van de problemen bij het "minen" van de literatuur is de ambiguïteit van termen zoals gen namen, en in dit geval in het bijzonder de ambiguïteit van de afkortingen of acroniemen van gen namen.
Het doel van hoofdstuk 4 is allereerst om de omvang van het ambiguiteitsprobleem te kwantificeren in de UniProt eiwit databank, en vervolgens een disambiguatie strategie te introduceren, gebaseerd op taxonomie. De hypothese is dat naarmate species taxonomisch meer verwant zijn, er een grotere kans is dat gen symbolen naar dezelfde biologische functie voor een specifiek gen verwijzen. Dit is gebaseerd op de aanname dat biologen en annotatoren vaak namen en functies kopiëren van gerelateerde soorten. Gebruikmakend van deze hypothese hebben we een algoritme ontwikkeld dat eenduidig een biologische functie kan toewijzen aan een (ambigu) gen symbool en zijn alternatieve namen, gebaseerd op de taxonomische relaties van de soort naam. De implementatie van het algoritme kan egbruikt worden om eenduidig en met hoge efficiëntie de lietratuur te doorzoeken, uitgaande van een gen symbool.

Tenslotte introduceert hoofdstuk 5 GeneIlluminator (GI), een applicatie die de disambiguatie methode van hoofdstuk 4 implementeert. De voorberekende sets van eenduidige synoniemen en hun biologische entiteiten zijn geïmplementeerd in een logisch (Boolean) of een vector model om PubMed abstracts met behulp van GoPubMed (een PubMed zoekmachine) op te halen. Gegeven een document doorzoekt GI alle categorieën onder welke dit document ondergebracht zou kunnen worden ("document-pivoted categorization"). Wordt een specifieke categorie opgegeven, dan doorzoekt GI alle documenten die onder deze categorie opgeslagen zouden worden ("category-pivoted categorization"). Hierdoor kan GeneIlluminator gebruikt worden om de abstracts uit de Medline database effectie te disambigueren.

Acknowledgement:

I am indebted to many people for their support and advices to the successful completion of my PhD degree and this dissertation.

My deepest gratitude goes to my supervisor Prof Jack Leunissen for believing in my capability of undergoing a PhD research position and for offering me this opportunity and his exceptional supervision. He has helped me during my four years doctoral research endeavour to move forward with investigation in-depth and to remain focus on achieving my goal. He has thought me how to methodically conduct research and has given me great role model. Very few professors still do the job themselves rather are administrative managers. I could not have made it so far without his endless inspirations.

Pieter Neerincx, should I say squash partner, you have been a real friend through out this period of ups and downs, I will never thank you enough, indeed you contributed substantially to the completion of this thesis by raising interesting issues and questions, always trying to explain everything even what I have done myself, not forgetting all the driving we made especially for my selfish interest, and your painting and moving skills they really helped. Thanks for making my social life in the Netherlands a positively memorable one, Once more, thank you mate. I will also thank Monique and Simon for being nice squash partners besides being infallible squash-lane booking reminder.

Harm, you have been a good buffer to the group, this involved your omni-assistance and care. I will also thanks you for being an excellent improvised couriers carrier, well done jobs, thanks to you my parent via Divine could receive on time original documents that facilitate their being here.

Ingrid Paffen, Thanks once more for assistance in filling in all the IND, Belastingdienst forms and striving to making my social life in the Netherlands unforgettable. Thanks once more for all the driving, you have been and still will be a very good "pilot".

Divine thanks for all the moral support and the memorable guided tours of Berlin, our friendship proved to go through years.

A special thanks goes to Jeroen Van Reeuwijk for acting as a mentor during my first months in the Netherlands and accepting of being a professional photographer on my defence ceremony

Special thanks goes to Peter Schaap and Peter Groenen for introducing me into bioinformatics the academic and company perspective of it respectively, all my warms thanks to all staff at the medical department and informatics at Organon, and Jan Polman my MSc. thesis supervisor in particular.

I will like to express my gratitude to all staff members at the laboratory of bioinformatics, Harm nijveen, Jifeng, Ernest, Judith, Hong, Linke, Arnold, Anand and all the former students and new students, it was casual and fun with you guys. A special "merci" to our shared-secretaries Maria and Mari-jose for their prompt and immediate services.

Thanks to members of the phytopathology group namely Bart Thomma, Peter, and Pierre. The cooperation we once had was valuable indeed. My appreciation also goes to the molecular biology group with whom occasional coffee break and birthday cakes were very fun. Special thanks go to Henk and Olga, for being always ready to listen and advise accordingly, I really appreciate.

I am grateful to my committee members, Prof. dr. G. Vriend, Dr. G. Jenster, Prof. dr. M. Groenen, Prof. dr. H.J.A. Bisseling for raising interesting issues and questions, for thorough reviews of my thesis report and for kindly accepting to be on my committee on a relatively short notice.

The literature discussion meeting have re-installed my belief in the benefits of well-focus discussion and here I would like to express my appreciation towards the group members, Roland Van Ham, Jack Leunissen and all the members from the applied bioinformatics group in plant research international (PIR) for their interesting discussion, and to the organiser of such inspiring meeting Roland and Jack.

This work would have been beyond imagination had it not been for the eternal support and encouragement and meticulous proofreading that I received from Jeannette Kluess.

I will like to thanks Prof. dr. H. Stunnenberg for believing in my capacity to fulfil a postdoctoral position in bioinformatics in his polyvalent group. Marc Van Driel for good introduction to my new position and enthusiasms toward sciences

Thanks to all friends I made out of the academic scope of this PhD, namely Barbara, Monique, simon, ottis and family, heidi, marsha and all those I forgot to mention, I appreciate your friendship.

I would like to express my appreciation to my father Etienne Alako, my mum Victoire Assongou, my sister Hortense Alako, my god father Christopher Ndeh, the Yamba family, Delphine, Gwendoline, Fanny, Wendy, Sandra for their love, support and encouragement.

I would like to express my sincere thanks to my wife liliane and my daughter Kaycee for their love and sacrifices.

This work was carried out in part using software provided by the Laboratory of Bioinformatics at Wageningen university and was sponsored by the Centre of BioSystems genomics (CBSG).

List of publications

Alako, B.T., Veldhoven, A., van Baal, S., Jelier, R., Verhoeven, S., Rullmann, T., Polman, J. and Jenster, G. (2005) CoPub Mapper: mining MEDLINE based on search term co-publication. BMC Bioinformatics, 6, 51

Alako, B.T., Rainey, D., Nijveen, H. and Leunissen, J.A. (2006) TreeDomViewer: a tool for the visualization of phylogeny and protein domain structure. Nucleic Acids Res, 34, W104-109.

Alako, B.T., Neerincx B.T., and Leunissen, J.A. (2008) A taxonomy-based disambiguation approach for gene names and symbols. (in preparation)

Alako, B.T., and Leunissen, J.A. (2008) GeneIlluminator: Disambiguation of PubMed Abstract (in preparation)

H. Peter van Esse, Emilie F. Fradin, Blaise T.F. Alako, Pierre J.G.M. de Wit and Bart P.H.J. Thomma (2008) Comparison of the transcriptomes of Cladosporium fulvum- and Verticillium dahliae-infected tomato (In preparation)

Lian-Feng Gu , Blaise T. F. Alako, Bing-Bing Wang, Zhen zhu (2008) Large-scale analysis of rice alternative splicing reveals their impact on protein domain and functions (In preparation)

About the author:

Blaise T.F Alako was born on 13 June 1976 in Sohock-nkondjock, Cameroon. After completing his basic education in French (école maternelle de Njigang-Nkondjock, école publique de Njigang-Nkondjock, école publique de Bomono-Ba-mbengue, lycée bilingue de Bonabéri-Douala, lycée polyvalent de bonabéri-Douala), he enrolled at the University of Buea-Cameroon with English as medium of education and received his BSc. Degree in Biochemistry in 1999 with first class honour. During his secondary and university education he won several scholarships for academic excellence. Between 1999 and 2001 his was giving private lessons to high school and university students as well as searching for sponsorship to further his education. In 2001 he was awarded a grant from the Netherlands Fellowship Programme (NFP) to study for his master's degree in Biotechnology at the Wageningen University and Research Centre. Keen to new ideas he dared enrolling for an MSc. Biotechnology/Bioinformatics instead. He obtained his MSc. Biotechnology/Bioinformatics from Wageningen University in January 2003. In November 2003 he was offered an AIO position in the Bioinformatics department of the Wageningen University and Research Centre. This thesis forms part of the outcome of this appointment. In November 2007, he was offered a Bioinformatics postdoctoral position in the Molecular Biology department at the Nijmegen Centre for Molecular Life Sciences (NCMLS) of the Radboud University.

Contact: alakotadon@yahoo.com
B.alako@ncmls.ru.nl

Education Statement of the Graduate School
Experimental Plant Sciences

The Graduate School EXPERIMENTAL PLANT SCIENCES

Issued to: Alako Tadontsop Francois Blaise
Date: 1 February 2008
Group: Bioinformatics, Wageningen University and Research Centre

1) Start-up phase	date
▶ **First presentation of your project**	
Arabidopsis genome wide survey for Leucine Rich Repeat (LRR)	Feb 17, 2004
▶ **Writing or rewriting a project proposal**	
▶ **Writing a review or book chapter**	
▶ **MSc courses**	
▶ **Laboratory use of isotopes**	
Subtotal Start-up Phase	*1.5 credits**

2) Scientific Exposure	date
▶ **EPS PhD student days**	
International PhD student day (SDV-EPS), Paris, France	May 19, 2005
EPS PhD student day (Wageningen)	Sep 19, 2006
EPS Phd Student day (Wageningen)	Sep 13, 2007
▶ **EPS theme symposia**	
EPS theme 4 symposium 'Genome Plasticity', Wageningen University	Dec 09, 2004
▶ **NWO Lunteren days and other National Platforms**	
▶ **Seminars (series), workshops and symposia**	
Center for BioSystems Genomic (CBSG) disease resitance workshop	Jun 10, 2004
Center for BioSystems and Genomics Intellectual Property (CBSG IPR) workshop	Oct 28, 2004
Center for Biosystem Genomic (CBSG) Bioinformatics Cluster meeting	Aug 27, 2004
Genomic momentum-Rotterdam	Aug-Sep 31-01, 2004
Working group Bioinformatics - Symposium 2004 (Hanzehogeschool Groningen)	Oct 07-08, 2004
EPS Symposium on system Biology (Pierre de wit tribute)	Nov 04, 2004
Prof. dr. Masahiro Yano (uncovering genetic control of flowering time in rice)	Jun 26, 2006
Wageningen Center for Food Science (WCFS) and Biology Information Technology (BioIT) meeting	Oct 17-18, 2006
Prof. dr. Scott Poethig (Regulation of phase change in plants by miRNAs and trans-acting siRNAs)	Sep 24, 2007
Benelux Bioinformatics Conference (BBC 2007) WICC-Wageningen	Nov 12-13, 2007
▶ **Seminar plus**	
▶ **International symposia and congresses**	
14th Annual International Conference on Intelligent Systems for Mol.Biol., Fortaleza, Brazil	Aug 04-05, 2006
Second Annual AB3C conference X-meeting, Fortaleza, Brazil	Aug 06-10, 2006
15th Annual ISMB-SIG 2007, Vienna, Austria	Jul 19-20, 2007
European Conference on Computational Biology (ECCB) 2007,Vienna, Austria	Jul 21-25, 2007
▶ **Presentations**	
Poster: Phenolink: A Data mining framework as catalyst for map based Cloning	May 19, 2005
Oral: Arabidopsis genome wide survey for Leucine Rich Repeat	May 19 2005
Oral: The marriage of clustering and taxonomic tagging at LSG meeting PRI	Jun 22, 2006
Oral and poster: TreeDomViewer (ISMB 2006 Software demo)	Aug 07, 2006
Oral: Visualization of protein Phylogeny and domain structure	Feb 22, 2007
Poster: Taxonomy-based disambiguation of all species gene names (ISMB Vienna 2007)	Jul 21-23, 2007
Oral: GeneIlluminator: Disambiguation of gene nomenclature and PubMed abstract	Sep 06, 2007
Oral: Disambiguation of PubMed abstract (Genomic Research 2007)	Nov 29, 2007
▶ **IAB Interview**	Sep 16, 2006
▶ **Excursions**	
Subtotal Scientific Exposure	*16.2 credits**

3) In-Depth Studies	date
▶ **EPS courses or other PhD courses**	
Wageningen Springschool Bioinformatics 2004, DATA triple: Information Integration, Interpretation	Mar 31- Apr 2, 2004
Course 'Comparative Genomics'	May 2006
Course 'Web services'	Sept-Oct 2007
Text Mining for Dutch Genomics	Nov 23, 2007
▶ **Journal club**	
Member of literature discussion group Bioinformatics	2004-2007
▶ **Individual research training**	
Auto-didact several prog.languages (Ajax, Perl, Python, Shell, Javascript, Java, MySQL, Awk, SVG)	2004-2007
Subtotal In-Depth Studies	*10.2 credits**

4) Personal development	date
▶ **Skill training courses**	
Dutch language course, level 1 (CENTA)	Jan-Sep 2004
▶ **Organisation of PhD students day, course or conference**	
▶ **Membership of Board, Committee or PhD council**	
Subtotal Personal Development	*2.2 credits**

TOTAL NUMBER OF CREDIT POINTS*	30,1

Herewith the Graduate School declares that the PhD candidate has complied with the educational requirements set by the Educational Committee of EPS which comprises of a minimum total of 30 credits

** A credit represents a normative study load of 28 hours of study*

Printing establishment:
Gildeprint drukkerijen BV, Enschede

Graphics on the cover:
Exclusively designed for this thesis by Blaise T.F. Alako