

## IDENTIFYING BIOMEDICAL CONCEPTS



July 13<sup>th</sup> 2006

## ANALYZING LITERATURE

On-going research processes have given way to relevant new discoveries, which in turn have triggered and enhanced the development of new therapies and new technologies. All acquired knowledge has been collected in scientific articles, describing working paths, methods and conclusions derived from each novelty achieved. Therefore, literature has become a significant sample of the scientific wisdom status.

In the case of Molecular Biology, available literature is a key element for the implementation of new basic research and the development of new medicines. Both technological advance and professional competition have contributed to so large a production of scientific articles that it is quite impossible for researchers to cover them all. The need of a more focused work and the pressure emerged from scientific community and pharmaceutical companies in order to obtain immediate results are two combined factors that have provoked the development of new information systems, which in fact are indispensable for research projects to success.

Throughout the years, new discoveries in the field of Molecular Biology haven't had a common formal element to define them. As new genes and molecular processes were discovered, researchers gave them a random name. That problem is especially obvious when it comes to genes' names; scientists have followed different procedures at the moment of baptizing their discoveries. Some genes have been named after non-related meanings (e.g.: Superman, jazz, etc.); others have their identity in relation with the developed activity (e.g.: tumor necrosis factor); there are cases where genes' names follow the nomenclature of related diseases (e.g.: FHM comes from family hemiplegic migraine disease and is also synonym of the gene generating the condition).

New initiatives have been born, together with scientific publications, in order to build collecting-oriented data bases on genes or proteins. Though they usually represent a reliable source of information, they are frequently incoherent when it comes to comparing contents with the general contents present in specialized literature. This situation is undeniable in examples such as the TNF, because data-bases' information contained in services like HUGO, Swissprot and Locus link is not sufficient, nor helpful towards automatic detection in published documents.

If aiming to analyze literature in the field of Molecular Biology, the knowledge of biology-based terms is required: names of genes and proteins, diseases, chemical compounds, etc. But, recognizing these elements or bio-entities is frequently difficult, mainly due to the following circumstances:

- Bioentities do not have a formal structure to be easily recognized by a computer system.
- As it occurs with terms of frequent use, one same word can refer to different things, depending on the context.
- Acronyms and symbols are largely used, thus multiplying the problem of ambiguity.

Synonyms	Homonyms	Acronym
Different word for the same biomedical entity	Same name for different biomedical entities	Reduce word representing a biomedical entity
<ul style="list-style-type: none"><li>• In Human there are at least 5.418 genes with synonyms (38% of the total genome)</li><li>• Drugs have a commercial name and a chemical name</li></ul>	Symbol PAP is an alias for: <ul style="list-style-type: none"><li>• PAP (Pancreatitis-associated protein)</li><li>• MRPS30 (Mitochond ribosomal prot 30s)</li><li>• PAPOLA (PolyA polimerase alpha)</li></ul>	SCT stands for: <ul style="list-style-type: none"><li>• Stem cell transplant</li><li>• Secretin</li><li>• Salmon calcitonin</li></ul>

The text is always subject to writing variations, and meaning is not always modified. Never the less, those variations can become a big problem for automatic systems (both IL 2 and IL-2 are valid terms when speaking about interleukin 2).

## THE BIOENTITY DETECTION SYSTEM OF AKS<sup>2</sup>

The bioentity detection system implemented in AKS analyzes scientific literature and retrieves biological meaning terms (bio-entities). The first step in the process consists in parsing and tokenizing the document and then to tag every word in the text, according to biological relevant categories. After that initial procedure the system continues with the analysis, based on different algorithms according to the type of bioentity to be analyzed

### Genes

Genes' names appear in two different forms in the scientific literature: as full names (a functional description of the gene, such as tumor necrosis factor" or "janus kinase"), and as gene symbols (an abbreviation or acronym, such as "TNF" or "JNK"). The approach for detecting both instances is different, since the associated problems also differ.

In the analysis of full genes' names, there are two phases; first the words are classified in different types (key words, types, location, stop words ...), based on rules

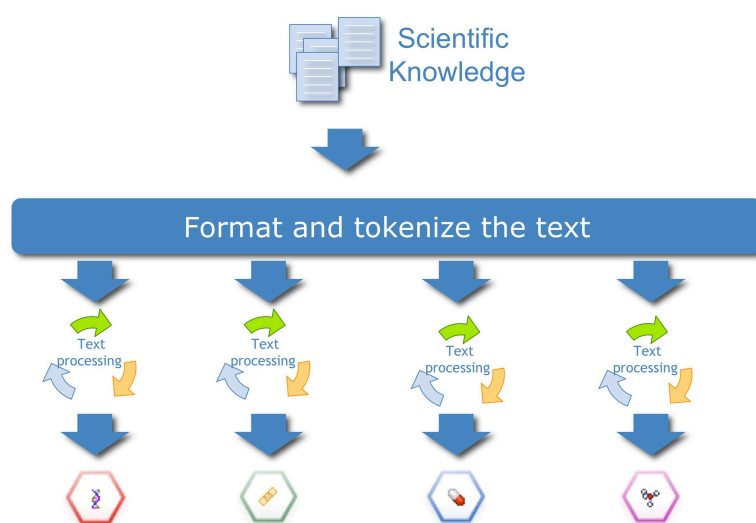
and pre-processed lists. In a second stage, sequences of type words relevant to determine whether the text is making reference to a gene or not are identified. Genes' names detected in this phase are treated as preliminary results.

Once the term is identified as a possible gene, it is compared to a system of predefined information. The comparison is made relying on previously established rules and scoring systems regarding word profiles. The preliminary results are then linked to a predefined biological meaningful term list, based on the score system in order to resolve ambiguities.

Symbol detection is a slightly more complex process, since the same symbol frequently represents different genes, or a disease, or even a non-relevant biological term. To distinguish whether a symbol refers to a gene, the system relies on the term classification described above, in predefined word lists and rules. The process follows two steps:

The system identifies possible gene symbols and assigns a probability of it being a gene name. This probability depends on:

1. The adjoining words of the possible symbol. These words are compared to a list of predefined terms associated to a different score, according to their relative position with regards to the symbol. There are terms such as "gene" or "protein" that increase that probability, and others such as "cell" or "plasmid" that reduce it. Many terms have scores that increase or decrease chances of success, always referring to the relative position next to the symbol.
2. Symbol structure. If the symbol can be confused with a normal word such as CAT, the probability of being a gene name decreases. The system relies as well on case sensitivity to affect the probability of being a gene symbol.



3. Neighboring words defined as genes. If any of the adjacent words around the symbol are defined as genes, the chances of the symbol being a gene name increase.

Once the symbol is defined, with certain reliability of being a gene name, the second step consists on recognizing which gene it refers to. In order to do so, the system follows the same analysis as the one described for full gene names. The difference is that, in this case, the system takes into account the symbol probability

### **Diseases and symptoms**

The dilemma behind the detection of diseases' names throughout the biomedical literature is similar to the one for genes' names. Actually, the internal analysis procedure is the same. The system identifies possible diseases' names starting from predefined rules and lists and resulting in preliminary diseases' results. These results are then compared to the information of external databases such as UMLS and OMIM.

In the diseases' search process, the system can find terms referring to general diseases such as diabetes or others more specific (e.g.: diabetes mellitus non-insulin-dependen).

### **Chemicals and drugs**

Chemicals compounds, and especially organic chemicals compounds, have a specific form or word morphology (e.g. glucose 6-phosphate, acetylcholine, n-methyl-d-aspartate, etc.) that can be recognized. These terms are found analyzing the morphology of the words. The detection of chemical compounds presents the following challenges:

1. There are words that, even if not naming a chemical compound, can follow these rules. The system uses dictionaries and contextual information to remove these false positives (e.g.: chloroplast, which contains the prefix "chlor", distinctive identifying element for chemical compounds).
2. There is a big variability in the use of signs such as dashes or parenthesis in the names of chemical compounds. There are cases in which a same chemical product can be found in the literature in 25 different ways (e.g.: (3,4,5-dimethylthiazol-2-yl)-2,5-diphenyl tetrazolium bromide) .
3. The chemical compound category is very heterogeneous, since it brings together entities such as chemical elements, cofactors and drugs, which in a high number of cases can not be detected by rules. The system relies as well on dictionaries to detect these types of entities.

The system bases its search for chemical compounds on predefined rules like algorithms and lists of positive result and/or exception conditions.

### **Bioterms and bioactions**

There are many biomedical relevant terms that might not feet to any of the above categories. Words such as "pregnancy", or concept such as "clinical trials" can help to focus the information search. The same concept can have different spelling forms due to grammatical form (activation, our activate), or just bye difference in american and british english (analyze, analyse).

The system groups the different spelling variants of these terms with the help of dictionaries and recognize them in the scientific text. Terms are categorized under bioterms by predefined rules, algorithms and positive and/or exception conditions.

In biomedicine processes are of grate interest to define new knowledge. Under the bioaction category there are key biomedical concepts that describe from molecular to physiological processes.

## ORGANIZING LITERATURE

The identification of the key concepts for biomedicine is a key step in the process of an in-depth and effective information extraction. This classification and organization of the text information will open a wide variety of possibilities:

- Perform conceptual searches. You can retrieve all the documents containing a certain term no matter how is written.
- Find key biomedical concept shared by a group of terms. You can better understand the clustering results of a DNA array experiment if you are able to see what each gene in a cluster have in common based on the literature.
- New discoveries. If a drug is related to a gene and this gene is related to a disease to which the drug is not related. Wouldn't it be interesting to try to explain that missing drug-disease relation?

The classification opens more possibilities however other steps on information management will need to be taken.