MILAN

A medical information-system linking agents to metabolic networks

B. Prins, R. Hofestädt

Bielefeld University, Faculty of Technology, Bioinformatics Department,

Universitätsstraße 25, D-33615 Bielefeld, Germany

bprins@techfak.uni-bielefeld.de, rhofestae@techfak.uni-bielefeld.de

Summary

The genome, the protein-biosynthesis and the metabolism constitute a complex, strongly linked system. The complete functionality of this system is not entirely understood yet. Especially pharmacologists and medical doctors are interested to know how the system works regarding the influence of medications (Drug-Pointing). In the last thirty years biochemical research delivered a huge amount of relevant data for this field. The problem is, that all this data is highly distributed in thousands of databases which are just partly interconnected. Gathering and interconnecting this large amount of distributed data by hand significantly slows down the research process. Hence it is vitally important to create tools in terms of information-systems, which support the process of understanding through automatic data-integration and analysis. In this paper we would like to introduce our approach of a corresponding information-system. MILAN is an extensible data-warehouse framework, which has already integrated about 7.7 million medical publication-abstracts as well as metabolic pathways and information about enzymes, ligands, agents and pharmaceuticals from a collection of major biochemical databases. The system provides a set of functionalities to show where the metabolism is affected by a certain drug and instantly displays the most relevant publications for the particular case. Future development will focus on the integration of more databases and the development of additional functions like detection of side-effects and drug-interactions.

1 Introduction

For drug-design and drug-enhancement the understanding of the metabolic system is essential. At present there are many unanswered questions. In particular, pharmacologists and medical doctors are interested in answering questions like: “How do specific medications affect the metabolism?” or “How can we avoid side-effects or drug-interactions?” To answer those questions a lot of data about all the metabolic system’s components is needed. This data is normally gathered through laboratory research and stored afterwards as electronic datasets in database-systems. In the past years more than thousand of those databases where established [1]. They provide data for all the different kinds of biochemical and molecular domains. For example, the KEGG [11] and REACTOME [12] databases contain data of metabolic pathways and their components, biochemical data about proteins are stored in PDB [7], ENZYME [18], UNIPROT [19] and others. All this information (and a lot more) is

important for an extensive observation of pharmaceutical effects in metabolic pathways. On one hand, the huge amount of information is an advantage; on the other hand, an efficient access to related but distributed datasets is difficult. Access to those molecular data stocks can be made easier by using data-integration techniques and the implementation of data-warehouses. Automization of the data-integration process is largely possible. So there are several other projects around, which concentrate on biochemical database integration. Projects like ENTREZ [15], COLUMBA [16], INTEG8 [17], to name just a few. Those projects have a broader thematic focus and do not specialize in drug-research. Our approach is to establish an extensible data-warehouse framework, which will focus on the influence of pharmaceuticals on the metabolism. MILAN automatically integrates and updates several biochemical databases as e.g. KEGG, PDB, RXLIST [13] as well as publication-databases like PUBMED/MEDLINE [20] and interconnects them. Within the data-warehouse structure a set of analytic algorithms is continuously implemented. Currently the MILAN prototype provides a variety of standard search methods and retrieval techniques, as well as functionalities to show where a drug affects the metabolism (Drug-Pointing). In addition, it simultaneously displays a list of the most relevant medical publications for the entered user-query. Future steps include the integration of more databases and the implementation of methods for detection of side-effects and drug-interactions. Together with an increasing number of functions, MILAN will constitute a complete working suite and could simplify the research of pharmacologists, medical doctors, and scientists. It also may assist in the development of pharmaceuticals in near future.

The paper is structured as follows: In the related works section we present other available solutions in this research field and describe how our system differs from them. In the following methods section we generate a simple theoretical model to map drug-related influences to a metabolic system. The complete system architecture, of MILAN will be described as well. The result chapter will show a typical use-case of MILAN: A search for the affected pathways and targets of the vitamin B1 derivate cocarboxylate. Finally we will give some conclusions and discuss the next steps in development.

2 Related works

In this chapter we will briefly describe other similar systems in our research field which have been implemented in the past.

MDDrugDB [10] – The Metabolic Disease Drug DB (University of Magdeburg, Germany)

This database focuses on the search for medications to treat metabolic diseases. It has no data-integration substructure. All data was entered by hand. Therefore the total amount of stored datasets is relatively low. No metabolic pathways are stored. Nor are any analytical procedures implemented.

PKB [21] – Pathway Knowledge Base (BMI Stanford, USA)

This system does not especially focus on the effect of drugs in metabolism, but automatically integrates pathway data and offers a lot of interesting search methods within the pathway data. PKB utilizes BIOPAX [23], an OWL-based standardized representation of pathway data, to integrate pathway information for disparate data sources.

Regarding the manner obtaining the required data we found several related systems which are integrating biochemical databases to generate surplus value. We can principally distinguish three main categories of integration projects, which can be split up into further subcategories:

- **Integration layers for biochemical databases**

  Universal integration-projects which integrate a set of biochemical databases in terms of an integration layer or wrapper. They mostly do not implement a user interface. They do not provide analytic algorithms working on the integrated data. The target group covers mainly developers of biochemical information-systems primarily located in the bioinformatics community.

  Examples are: MMIS [14] (standalone), BIODATASERVER [24], BIOWAREHOUSE [25], etc.

- **Universal biochemical multi-database-systems**

  These are Databases which integrate or wrap a set of biochemical databases. They interconnect the datasets of the different sub-databases and display a own schema. Mostly there are graphical user interfaces available. In recent times the GUIs appear as web interfaces. Some of the systems provide analytic algorithms working on the integrated data. Furthermore some of them use analytic functions of the root database-system or other sources. Those systems find exertion in many fields of biological research.

  Examples are: ENTREZ [15], COLUMBA [16], INTEG8 [17], PKB [21], etc.

- **Specialized biochemical or medical data-warehouses**

  These systems focus on a very specific theme and target-group. They rely on a defined dataset, integrated in a materialized manner from special biochemical databases. They do not map the sub-databases completely, but concentrate on the data which is important for a given context. They also restructure the data for their demands and provide special features suitable for the target-group. MILAN can also be classified into this group.

  *Examples are: BIOPATH/C@ROL [27], MMIS/*METABOLIKA [14], *PROTON [26], etc.*

3 Methods

At first we like to recollect some basics of pharmacology and then construct the underlying biological model for our application.

3.1 Underlying model

Drugs or agents are biologically active substances or mixtures of these substances. In an appropriate application they should heal or prevent trouble and diseases of the human and animal body. They substitute missing agents which are normally produced in the body, they suppress or destroy pathogens and some drugs are even able to produce psychological effects [2]. All drugs have in common to have some impact on specific metabolic pathways. (See also figure 1.) This influence strongly depends on the application, distribution, decomposition, disposition or expulsion of the bioactive agents (pharmacokinetics) and their special biochemical effect (pharmacodynamics) [3].

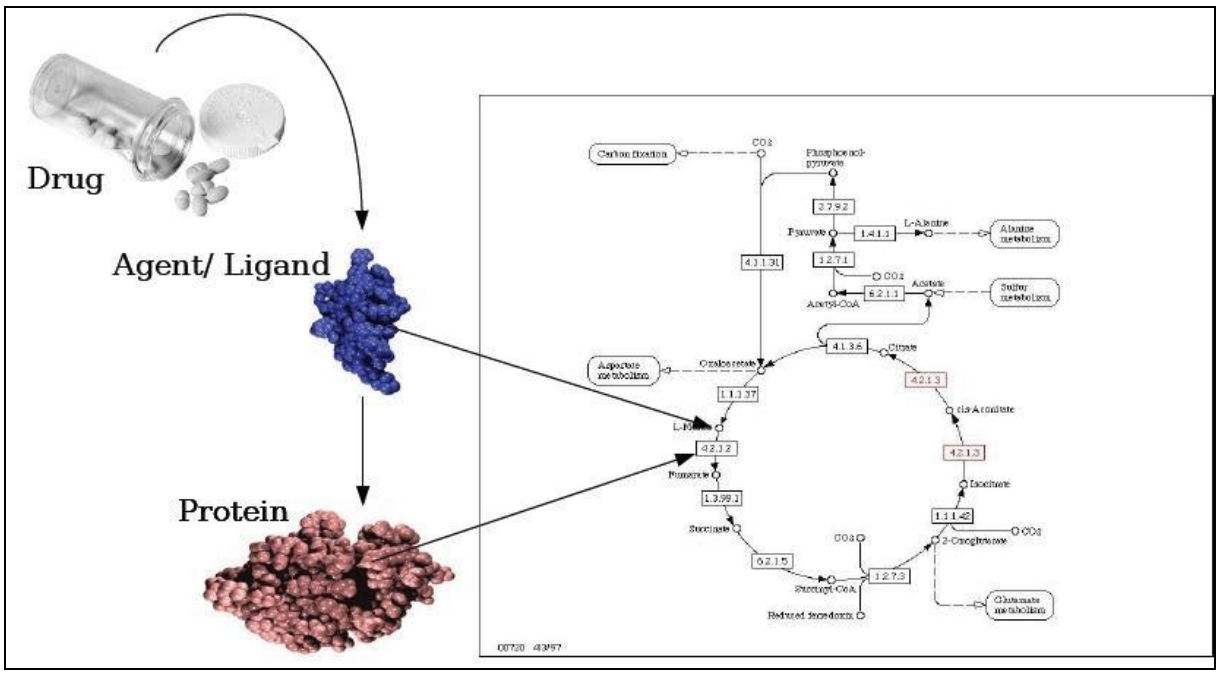

Figure 1: The figure shows the relation between medications and the metabolism. Drugs and their ingredients have a specific counterpart, mostly a protein. Agents and proteins are part of one or more specific metabolic pathways. Finding the location of medication-affected enzymes and receptors is also called drug-pointing.

We can imagine the central working point of a drug in a simple model (See also figure 2.). Every agent has a specific counterpart, a so called target-molecule. Most agents are small organic molecules, whereas most target-molecules appear as proteins. The interaction between agent and target is pretty well explained trough Emil Fischer's Key-Lock-Principle [4] and the theory of “induced fit” posited by Koshland [5] in the late 1950s. The simplest case is that one agent interacts with one specific target-molecule. After docking to the target-molecule, channels will open or close, or the production of substances will be activated or deactivated. This in turn initiates a cascade of further effects and will end up in a specific physiological effect.

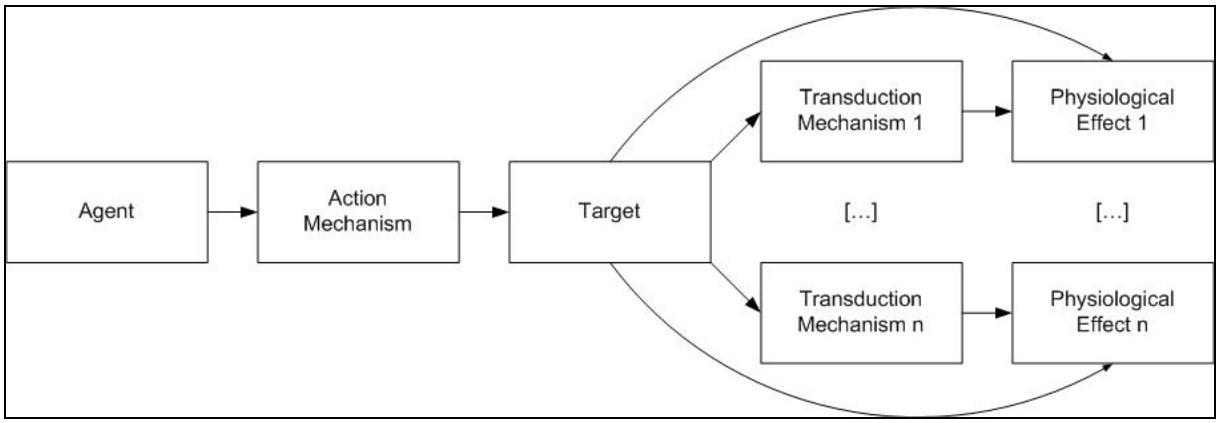

Figure 2: Model for agent-target interactions and their resulting physiological effects.

Of course there are a few exceptions that are not explained by this simple model. For example it is possible an agent interacts with more than one target-molecule, or it is even possible that many different ligands interact with the same target. Our model focuses on questions in pharmacodynamics but does not care about pharmacokinetics and pharmacogenetics. Nor is the field of genetic therapy handled.

With this given model it is easy to understand what kind of data we need for our system. Most important is data about metabolic pathways and all molecules involved. Those molecules are proteins (enzymes and receptors) and their ligands. The majority of drugs or agents appear as ligands, so we need information about drugs as well and link them with ligand data. To provide recent research results a sufficiently large set of publications is needed, too.

3.2 Implementation environment of MILAN

The MILAN base system and the user interface are implemented in the Java/J2EE programming language. The system is a purely web based system and needs no installation. All services run on a Linux system and they communicate with the client via an Apache Tomcat web-server. For operation the user just needs a standard web-browser. This architecture facilitates to use any operating system with an included internet browser. To save development time we also rely on a few third party libraries like OVERLIB [28] for tool-tips, JFREECHART [29] for statistical chart generation, and the KEGGAPI [11] for generating marked pathways.

3.3 Architecture of MILAN

MILAN can be classified as a medical data-warehouse for decision support. Any data-warehouse should comply at least with the following standards [8], [9]:

- The data warehouse has to make information easily accessible.
- The contents of the data warehouse must be understandable.
- The data warehouse must present information consistently.
- Data must be carefully assembled from a variety of sources, cleansed, quality assured.
- The data warehouse must be adaptive and resilient to change.

- There are analytic or simulating functions or tools implemented which work on the integrated data.
- The tools that access the data warehouse must be simple and easy to use.
- A query should produce minimal wait times.
- The existing data and applications should not be changed or disrupted when new data is added to the warehouse.

MILAN fulfils these standards as far as possible. It is easy to use, consistent, automatically updated and is a modular system which can be easily extended with further modules. The structure of MILAN is divided into three parts (See also figure 3). The frontend, the backend and the actual database management system. The frontend system is a J2EE web application processing the user queries. It communicates via the webserver with the database management system and displays all results within a webpage (See also figure 5). The backend-system controls the whole data integration process as well as the update and fallback management. All data is stored in a MySQL [30] database management system.

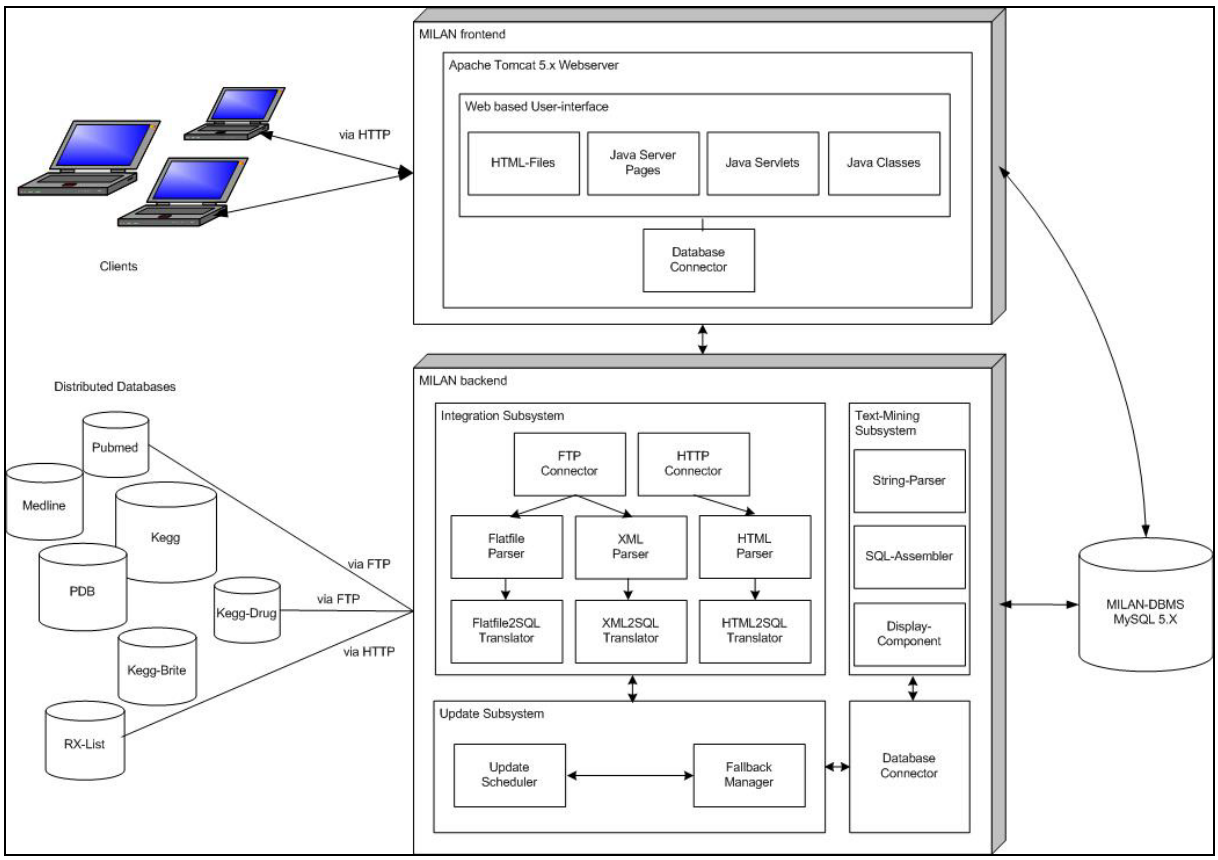

Figure 3: The figure shows the complete MILAN system structure. MILAN consists of three parts. The core system functionality is provided through the backend-system. It contains three subsystems which integrate all required data, control the updating process, react on integration failures and implement the textmining feature. The MILAN-frontend provides the web-based user GUI which processes the user queries and sends them as SQL-statements to the database. The third part is the actual database management-system which contains all gathered data.

3.4 Data integration & data-sources

The whole included data is automatically integrated via an extensive integration subsystem. The data needed for our system was integrated from freely available databases:

- KEGG
  metabolic pathway data / enzymes / ligands
- KEGG-DRUG
  drug / agent data
- KEGG-BRITE
  disease data
- PUBMED / MEDLINE
  medical/ biochemical publications
- RXLIST
  drug data like brand names and drug descriptions
- PDB
  3d- structural protein data

Because of the different types of interfaces of each database we had to implement different integration layers. Most of the data is integrated via an ftp connection and parsing of flat-files or XML-files. Afterwards the data is stored into the local MySQL 5.0 database management system. The complete dataset is reorganized and newly interconnected for MILAN and will be updated daily by an update subsystem which keeps all data up to date. A fallback subsystem takes care of the all-time-availability of the system, even when the update process fails and corrupt or incomplete data is delivered (See also figure 3.).

3.5 Drug-Pointing component

MILAN contains mostly classical agents. It provides searching methods for known drugs and their agents and yields a set of related metabolic pathways and the exact agent-target position. In addition, current scientific results in terms of publication-abstracts will be shown simultaneously.

The user can enter a name, CAS-number or group of an agent he or she is interested in. The system will deliver a result list of one or more agents. After selection of one agent the system searches for metabolic pathways in which the agent is involved. Via a SOAP web-service of the KEGG database, MILAN creates one ore more new pathway maps including a yellow marked agent. The system also looks up the publication database for related publications and shows the results sorted by relevance on the right side of the screen. A typical case is shown in the section results.

Additionally, MILAN offers a search function for diseases in which the user is able to look up which metabolic pathway is the root cause of the disease. In further development this search function will also show agents that are known to have effect in those particular pathways. Similar to the standard drug-pointing search, a pathway-map will be generated and all effecting agents will be labelled.

3.6 Text-mining component

Another important feature is the publication search. For any entered query MILAN does not only deliver a result-set of molecules, diseases or pathways, but also a relevancy-ranked list of related publications for the current subject. Even a standard publication search is implemented. To enhance the search as we know it from PUBMED we developed a two-phase searching method. First the user enters a string he or she is interested in. Afterwards a ranked result-set is shown. Now the user selects one abstract which fulfils his demand and the system will display this abstract. Until now the process is the same as we know it from several other publication databases. While showing the abstract, MILAN will search for all words in the title of the current abstract within the abstract-database, and displays the most relevant entries simultaneously. With this option the search will be refined automatically and remains in the given context (See also figure 4.). Because of the large amount of stored abstracts, this simple method delivers very good results. MILAN currently searches for entered strings in about 7.7 million abstracts from the year 1900 up to today. The data-size of the integrated publication-data is about 11 gigabytes. To do a fulltext-search in such a huge amount of data we need extreme efficient access structures, because time is the limiting factor to make this feature useful. Furthermore we need a meaningful ranking method to find only the most relevant abstracts.

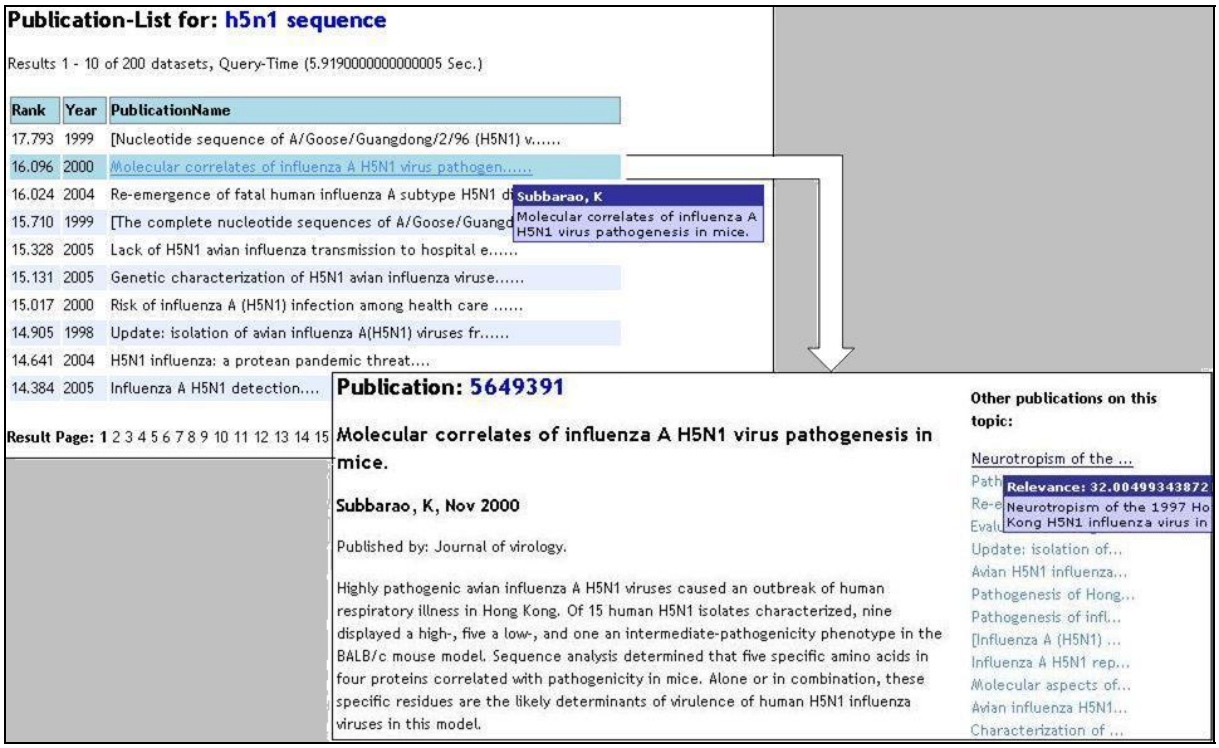

Figure 4: Here details of the MILAN-frontend are shown. On the upper left side we see a result set for the search:"h5n1 AND sequence". The user selected the second abstract from this resultset and gets the selected paper-abstract and a relevance-list for related papers on the right side. This list is generated by a background search for all words (except if they are in the stopword-list) of the title in the current abstract. In this case the search would be: "Molecular AND correlates AND influenza AND H5N1 AND virus AND pathogenesis AND mice". The words "of" and "in" are in the stopword-list and have been eliminated.

To solve both problems, we rely on a MySQL fulltext-index with ranking option and compression abilities of the local database [6]. This functionality is widely used in fulltext search of web based database systems and even in commercial web portals.

To speed up the search and eliminate very irrelevant terms we defined our own stop-wordlist. A stop-wordlist contains all words which are not needed for a search, like “and”, “the” or “is”, etc.. Those terms will inevitably have a relevance weight of zero. So if the entered string contains a word of the stop-wordlist the word will be ignored. To rank the result set for relevancy the datasets will be weighted. MySQL offers us the functionality to do that relatively easy. Three weights will be calculated: The local weight, the global weight and the query weight. The local weight measures how often a term appears in a row. So if a term appears many times in a row, the weight is high. The global weight depends on an inverse multiplier. So if a term appears in many rows of the result set, the weight is low. This is because a term which appears many times might not be as relevant as terms which appear seldom. To prevent the system from hanging up or delivering irrelevant results there is another feature built in. If over 50 percent of all rows in a table will contain a certain term, the fulltext-search will stop and deliver no results. Finally it will be measured how often the term is contained in the query (Qf). Now we can calculate the actual weight of a row, the query weight. It is calculated as follows:

$$QueryWeight = LocalWeight * GlobalWeight * Qf$$

4 Results

In this chapter we will show a typical use-case within MILAN. We will search for affected pathways by the drug cocarboxylase. In addition we will try to find related target proteins as well as publicised research results about the effect of this drug.

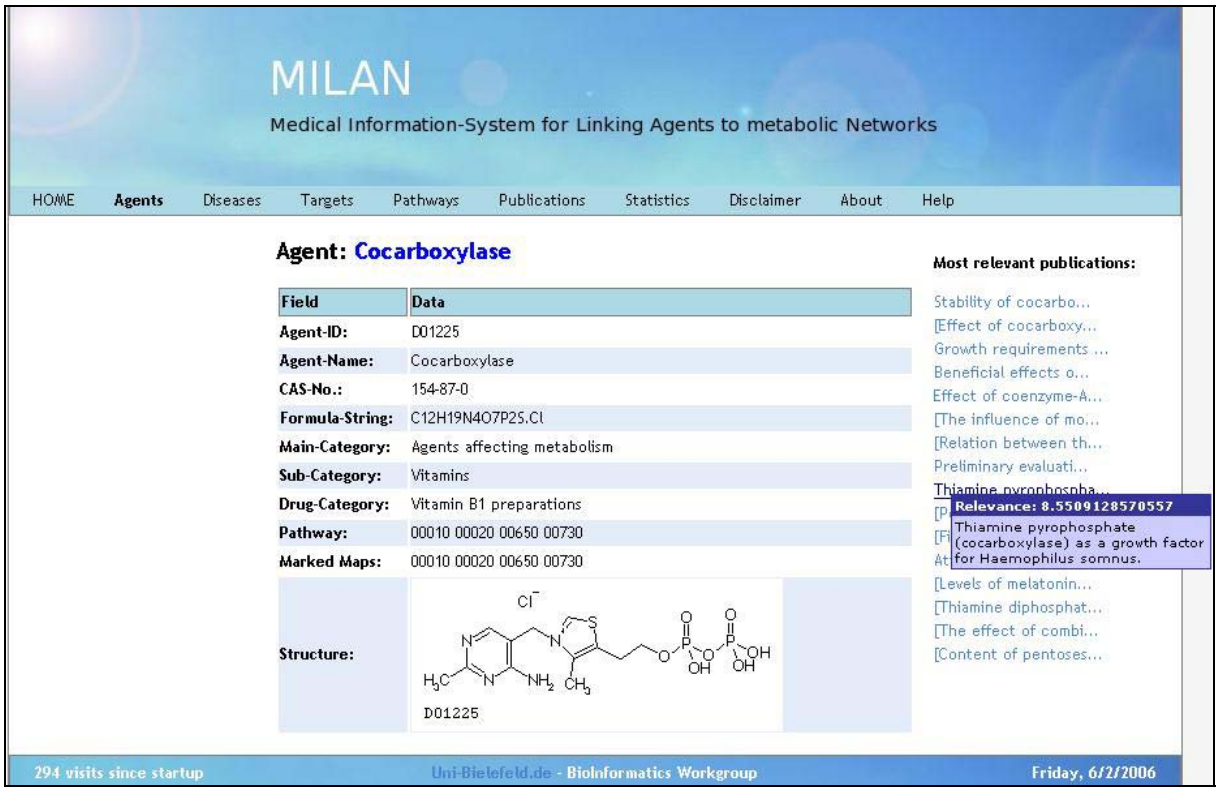

Figure 5: The figure shows the MILAN-webfrontend while displaying the dataset for a selected agent (cocarboxylase). In the middle of the screen we see information about the agent and links to marked pathway maps as well. By clicking these links the user will see a generated pathway-map with the highlighted agent-molecule (Figure 6). On the left side a list of the most relevant papers about this agent is shown.

Cocarboxylase or thiaminpyrophosphate or for short ThPP, is a vitamin B_1 derivate which is essential for the carbohydrate metabolism [22]. Lack of this substance may induce for e.g. diseases like beriberi.

After entering the term cocarboxylase in the agent search within the MILAN frontend, the system will show information about this molecule (See also figure 5). Containing CAS-number, formula-string, drug-categories and all related pathways in which cocarboxylase plays a role. MILAN now generates via a web-service labelled pathway maps of the related metabolic pathways (See also figure 6). They are available via the links under marked maps (See also figure 5). In these marked maps we can see the labelled agents and all target molecules of this specific metabolic pathway.

On the right side the system provides the 15 most relevant publications for the term cocarboxylase ordered by the query weight. (See also chapter 3.6). If these 15 publications are not enough, the user can use the separate publication search under the publications tab. There it is possible to browse through a lot more publication-abstracts. Entering conjunctive search strings is possible, too.

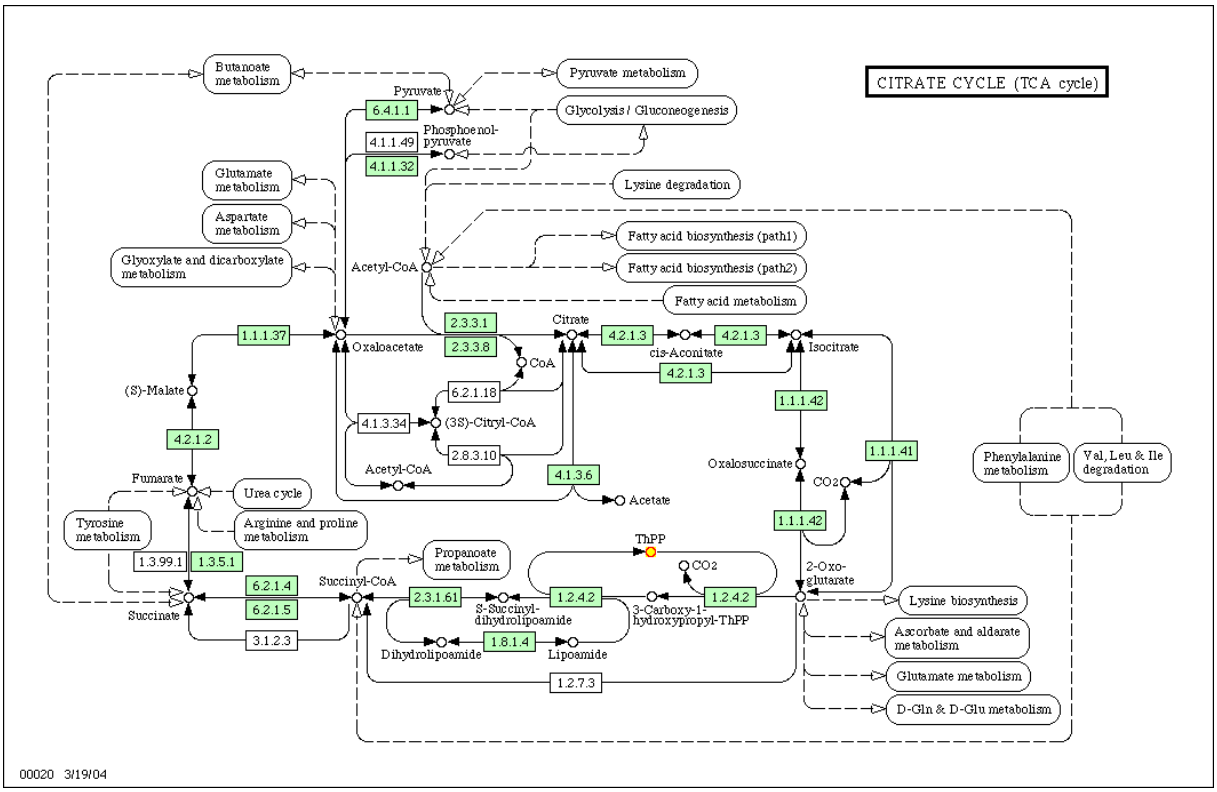

Figure 6: The figure shows a generated marked pathway-map for the selected agent Cocarboxylase which will be displayed after selecting one of the marked-maps links (see also figure 5). The example shows the human citrate cycle with the cocarboxylase (ThPP) reactant marked in yellow.

5 Discussion

MILAN is a prototype of an extensible data-warehouse system which describes how pharmaceuticals or just agents work within the human metabolism. As any data-warehouse MILAN follows the concept “all in one place”. So the user is able to investigate specific reaction-chains and drug-effects without looking them up separately in a large amount of distributed databases. MILAN integrates and reorganizes the distributed data and provides them together with recent scientific experimental results in terms of publication-abstracts. This combination should help to get new cognitions in this specific area.

Current state:

At present MILAN contains 2854 datasets about agents, 14099 datasets of chemical compounds, 4579 datasets for enzymes, as well as 180 metabolic pathways and 164 infectious diseases and metabolic disorders.

Additionally, the system allows to search in over 7.7 million paper-abstracts for relevant scientific-results in a selected area. Through the use of fulltext-indices a logarithmic time complexity for access could be reached. With a relevance-ranking mechanism based on a calculated real number it is possible to order the result sets by relevance. Another helpful feature is a context search that offers other directly related papers for any selected paper. Even during a search for an agent-target relation a result-set of relevant papers will be shown at the same time. The complete database has a size of 12 gigabyte now and gains daily.

The MILAN prototype is available via http://bprins.homelinux.net/MILAN-BETA.

Future development:

Currently we are expanding MILAN in this direction:

- Displaying more information about drugs (brand names, descriptions, etc.).
- Providing information about all effecting drugs within a disease related pathway.
- Showing the PDB protein 3-D-models within a viewer like JMOL [31].
- Integrating more relevant databases.
- Providing more analytic / statistical functionality.

A recent survey of the German Bundesapothekerkammer showed that in Germany almost seven thousand prescriptions where faulty, because of wrong diagnoses or unregarded drug-interactions. Therefore, the next steps will be the integration of new modules, like components which help to find new hints about unknown side-effects, and drug-interactions. Also a module for virtual screening is planned to test agent-target combinations for docking-affinity.

6 References

[1] S. Suhai. Theoretical and Computational Methods in Genome Research. New York: Plenum Press, 1996.

[2] O. Strubelt. Elementare Pharmakologie und Toxikologie (4th Edition.). Stuttgart: Fischer UTB, 1991.

[3] G. Pöch, H. Juan. Wirkungen von Pharmaka (2nd Edition.). Stuttgart, New York: Thieme, 1990.

[4] E. Fischer. Einfluss der Konfiguration auf die Wirkung der Enzyme. *Berlin: Deutsche Chemische Gesellschaft* 27: 2984–2993, 1894.

[5] D. Jr. Koshland. Application of a theory of enzyme specificity to protein synthesis. Proc. Natl. Acad. Sci. USA 44: 98–104, 1958.

[6] P. DuBois. MySQL - The definitive guide to using, programming and administering MySQL 4.1 and 5.0, Third Edition. Indianapolis USA: Developers Library, 2005.

[7] RSBC PDB. The Protein Database, www.rcsb.org/pdb, New Jersey, USA: Rutgers State University of New Jersey

[8] W.H. Immon. Building the Data Warehouse, 3rd Edition. New York: Wiley & Sons, 2002.

[9] R. Kimball, M. Ross. The Data Warehouse Toolkit, 2nd Edition. New York: Wiley & Sons, 2002.

[10] MDDRUGDB. The Metabolic Disease Drug Database, http://mddrugdb.genophen.de/

[11] KEGG. Kyoto Encyclopedia of Genes and Genomes, www.genome.jp/kegg/

[12] REACTOME. A curated knowledgebase of biological pathways, www.reactome.org/

[13] RXLIST. The Internet Drug Index , http://www.rxlist.com/

[14] MMIS/METABOLIKA: A. Freier, M. Hoding, R. Hofestädt, M. Lange, U. Scholz. Molecular database integration: Analysis of metabolic network control

[15] ENTREZ. The life sciences search engine, www.ncbi.nlm.nih.gov/gquery/gquery.fcgi

[16] COLUMBA. Protein structure annotation database, www.columba-db.de/

[17] INTEGR8. Integrated genome and proteome browser, www.ebi.ac.uk/integr8/

[18] ENZYME. Enzyme nomenclature database, www.expasy.org/enzyme/

[19] UNIPROT. Universal Protein Resource, www.ebi.ac.uk/uniprot/

[20] PUBMED/MEDLINE. Publication database, http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?DB=pubmed

[21] PKB. Pathway Knowledge Base, http://pkb.stanford.edu/

[22] A.L. Lehninger, D.L. Nelson, M.M. Cox. Prinzipien der Biochemie, 2. Auflage. Heidelberg, Berlin, Oxford: Spektrum Akademischer Verlag, 1994.

[23] BIOPAX. Exchange format for biological pathway data, www.biopax.org

[24] A. Freier, R. Hofestädt, M. Lange, U. Scholz, A. Stephanik. BIODATASERVER: A SQL-Based Service for the Online Integration of Life Science Data, In Silico Biology, 2(2), 2002.

[25] BIOWAREHOUSE: Database integration for bioinformatics, http://bioinformatics.ai.sri.com/biowarehouse/

[26] PROTON: Prototyping Object Networks, http://tunicata.techfak.uni-bielefeld.de/proton/

[27] BIOPATH/C@ROL: Biochemical pathways database, www.mol-net.com/biopath/

[28] OVERLIB: Javascript tooltip library, www.bosrup.com/web/overlib

[29] JFREECHART: Java chart library, www.jfree.org/jfreechart/

[30] MYSQL: Database management system, www.mysql.com

[31] JMOL: Java 3-D PDB Viewer component, www.jfree.org/jfreechart/