

Bachelor Thesis

ERIS - Enzyme Reaction Information System

Franz Fenninger



Janet Thornton Group
EMBL European Bioinformatics Institute
Wellcome Trust Genome Campus
Hinxton, Cambridge, CB10 1SD, United Kingdom



Institute for Genomics and Bioinformatics
Graz University of Technology
Petersgasse 14
8010 Graz, Austria

Supervisor EBI:
Dr. Syed Asad Rahman

Supervisor TU Graz:
Univ.-Prof. Dipl.-Ing. Dr.techn. Zlatko Trajanoski

Graz, June 21, 2009

Abstract

English

ERIS is a free Java software for chemists to examine molecules and enzymatic reactions. Besides visualization and editing of chemical structures it also provides interaction with a database storing small molecules and enzymatic reactions. Furthermore it uses a couple of chemoinformatics libraries enabling the user to conduct Similarity - as well as Sub- and Superstructure Searches for molecules. But also reaction searches based on the compounds structure or the bond changes can be performed.

In order to understand how a reaction takes place it is possible to execute an Atom-Atom-Mapping algorithm which can be utilized to visualize the bond changes taken place and also the transition state of that reaction.

Since the database not only stores the structure but also physico-chemical properties of molecules it becomes possible to search for physico-chemical properties yielding appropriate molecules.

German

ERIS ist eine frei verfügbare Java software für Chemiker um Moleküle und enzymatische Reaktionen zu untersuchen. Neben der Visualisierung und Bearbeitung von chemischen Strukturen ist es auch möglich auf eine Datenbank welche Moleküle und Enzymreaktionen speichert zuzugreifen. Es werden verschiedene Chemoinformatik Bibliotheken verwendet, um Ähnlichkeits-, Sub- und Superstrukturen Suchen für Moleküle zur Verfügung zu stellen. Weiters ist es aber auch möglich Reaktions-Suchen basierend auf deren Edukt- und Produktstrukturen oder deren Bindungsänderungen durchzuführen.

Um den Ablauf einer Reaktion besser zu verstehen ist es möglich einen Atom-Atom-Mapping Algorithmus anzuwenden mit dem in weiterer Folge die Bindungsänderungen einer Reaktion sowie deren bergangszustand angezeigt werden kann.

Da in der verwendeten Datenbank nicht nur die Struktur sondern auch physikalisch-chemische Eigenschaften der Moleküle gespeichert ist, kann auch nach diesen Attributen gesucht werden und passende Moleküle dafür ausgegeben werden.

Contents

1	Introduction	2
2	Background	3
2.1	Used Software and Libraries	3
2.1.1	Chemsitry Development Kit	3
2.1.2	JChemPaint	3
2.1.3	Jmol	3
2.1.4	Chemlib	4
3	Methods	5
3.1	General Functions Added	5
3.2	Atom Atom Mapping	5
3.3	Dugundji-Ugi-Model	6
3.4	Reaction Mechanism Database	7
3.4.1	Database wrapper	9
3.5	Searches	9
3.5.1	Fingerprints	9
3.5.2	Maximum Common Subgraph	10
3.5.3	Physicochemical Properties	10
4	Results	12
5	Discussion	16
5.1	Future improvements	16
6	Acknowledgements	17

Chapter 1

Introduction

ERIS is a free Java software which was developed at the EMBL European Bioinformatics Institute in Prof. Thorntons Group. The aim of the software is to give a better understanding of chemical reactions and to perform different molecule and reaction searches. Therefore the target audience are chemists. For example one major task of chemists is to find or create compounds with desired properties. In order to find the appropriate chemical structures a structure-property relation needs to be established. This relationship is called Quantitative Structure-Activity Relationship (QSAR). Provided with the requested properties for a chemical structure ERIS can search the ReactMechDB for possible matches. Consequently these molecules can then also be used again to search for reactions capable of synthesizing this structures.

The project team during the main developing phase in summer 2008 consisted of my Supervisor Dr. Syed Asad Rahman who also developed most of the chemical algorithms used by the software, Dr. Lorenzo Baldacci who was responsible for the database while my part was to develop a front end utilizing the chemical algorithms and interacting with the database.

Therefore this paper mainly describes this front-end but also gives some information about the database and the algorithms it is based on and the interaction between these parts.

Chapter 2

Background

2.1 Used Software and Libraries

2.1.1 Chemsitry Development Kit

The CDK is an open-source Java library for the use in bio- and chemoinformatics. It provides functions for 2D diagram editing and generation, Quantitative Structure-Activity Relationship (QSAR) calculations of molecules as well as the export to several chemical file formats. It was developed in 2000 by Christoph Steinbeck, Egon Willighagen and Dan Gezelter and is used in several applications like JChemPaint or Bioclipse and also builds the basis for the ERIS.

2.1.2 JChemPaint

JCP is an open source software written in Java for drawing 2D chemical structures and was created by Christoph Steinbeck and Stefan Krause. It now uses the CDK library and provides functions for drawing bonds and atoms to create a molecule and subsequently reactions. These chemical structures can be layouted with a cleanup functions and saved to several file formats. In ERIS a lot of JChemPaint functions are used to create and visualize molecules in 2D.

2.1.3 Jmol

Jmol is the 3D analog to JCP, an open-source Java viewer for chemical structures in 3D. It provides several visualization possibilites like the ball-and-stick model and a labelling option for all atoms. Chemical structures can be resized and rotated in a 3D space.

2.1.4 Chemlib

The Java library Chemlib was developed by Dr. Syed Asad Rahman and Dr. Lorenzo Baldacci at the EBI and provides another broad selection of functions for chemoinformatics. It is also based on the CDK and extends and modifies several CDK classes for the use in ERIS. Algorithms like the Atom Atom Mapping for reactions or the creation of the R-Matrix which are crucial for the ERIS software are part of this library.

Chapter 3

Methods

3.1 General Functions Added

Contrary to the JChemPaint software which can only open one chemical structure in each instance ERIS was based on a Multiple Document Interface so it can handle several chemical structures at once. Therefore a Drag And Drop support was also added to the program so it became possible to for example move molecules from one document to a reaction in another document and declare them as either reactants or compounds in this reaction. This facilitates the drawing of reactions significantly.

ERIS is also able to retrieve chemical structures from a database which holds molecules and reactions from renowned databases like KEGG, Chebi or IntEnz or Macie. The structures, retrieved from the database can also be integrated easily in a new reaction.

Using fingerprints ERIS can also perform structure- and similarity searches for both reactions and molecules. Furthermore it is also possible to search for molecules by physicochemical properties. With the functions provided by the Chemlib library ERIS is able to create the Atom Atom Mapping as well the R-Matrix of a reaction.

Finally using Jmol chemical structures with the appropriate coordinates can also be displayed in 3D.

3.2 Atom Atom Mapping

The purpose of Atom Atom Mapping is to identify which atom on the reactant side of an enzymatic reaction is transferred to which product atom. Therefore the reactant- and the product-atoms each get a unique labelling (Figure ??).

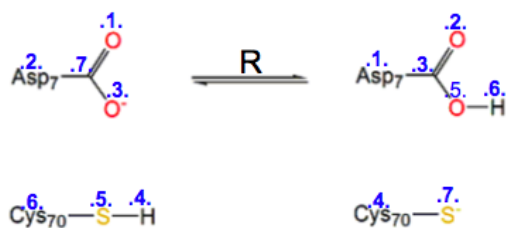


Figure 3.1: Labeled reaction

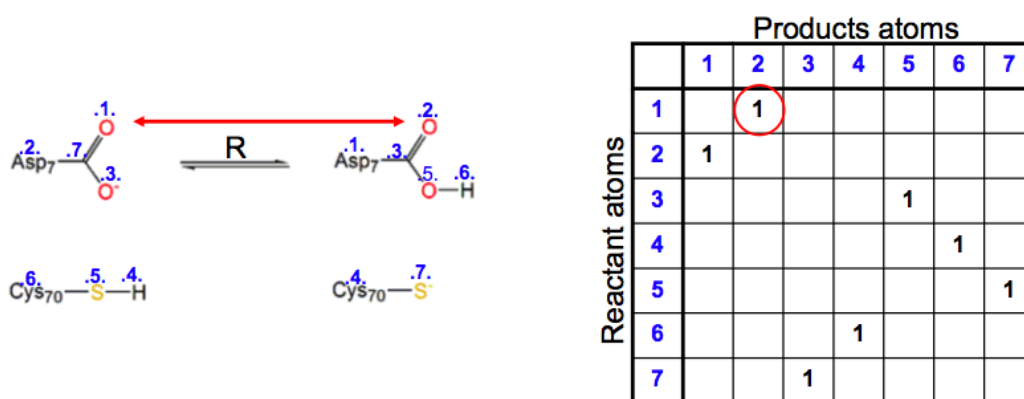


Figure 3.2: Mapped atom represented in a matrix

The mapping is then represented by a matrix whose rows represent the reactant- and columns the product-atoms. At the intersection of a row and a column whose atoms are mapped a flag stands representative for the mapping (Figure ??).

In order to calculate the mapping of a given reaction a MCS (Maximum Common Subgraph) algorithm is used. To facilitate the understanding of how a reaction is mapped in the ERIS atoms on the product side of a reaction are always displayed with the same ID as the mapped reactant-atom, so that two mapped atoms always share one ID.

3.3 Dugundji-Ugi-Model

The DU Model describes reactions through mathematical operations in the form of reaction matrices (R-Matrices). These R-Matrices represent the electron-transfers and the bond changes taking place in a reaction.

In the DU Model the reactants and products are represented by a bond-electron matrices (BE-Matrices). The rows and columns of the matrix stand

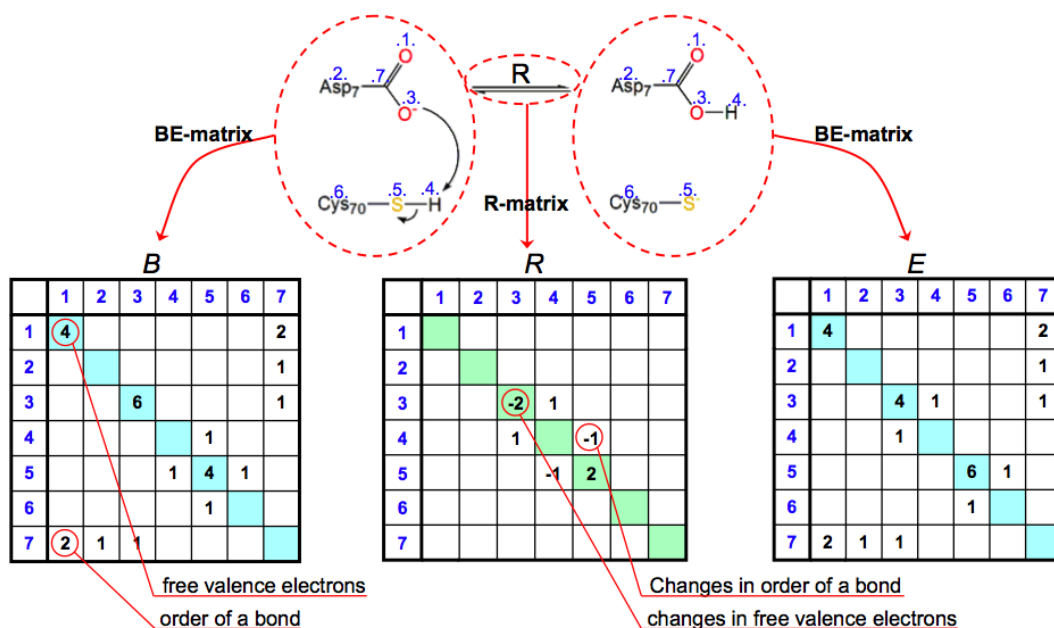


Figure 3.3: DU Model

for the atoms of the molecule. The diagonal elements store the number of free valence electrons of the corresponding atom and in the off-diagonal elements information about the connectivity between two atoms is held. Therefore a 1 represents a single-bond, 2 a double-bond and 3 a triple-bond. A reaction can now be expressed by a mathematical equation.

$$B + R = E \quad (3.1)$$

Here the BE-Matrix B represents the educt molecules while E stands for the molecules on the product side. The R-Matrix describes the changes between these two Matrices and therefore contains information about which molecules are cleaved and formed in a reaction. In order to calculate the matrices of the DU-Model the atoms of the reactant and product side have to be mapped to each other. This is warranted by the Atom Atom Mapping algorithm.

3.4 Reaction Mechanism Database

The main purpose of the database is to store the structure of small molecules. For each molecule all the atoms and connecting bonds including their order and stereo-type are saved. Also reactions are stored by referencing to the compounds it consists of and saving the reaction direction. Furthermore there

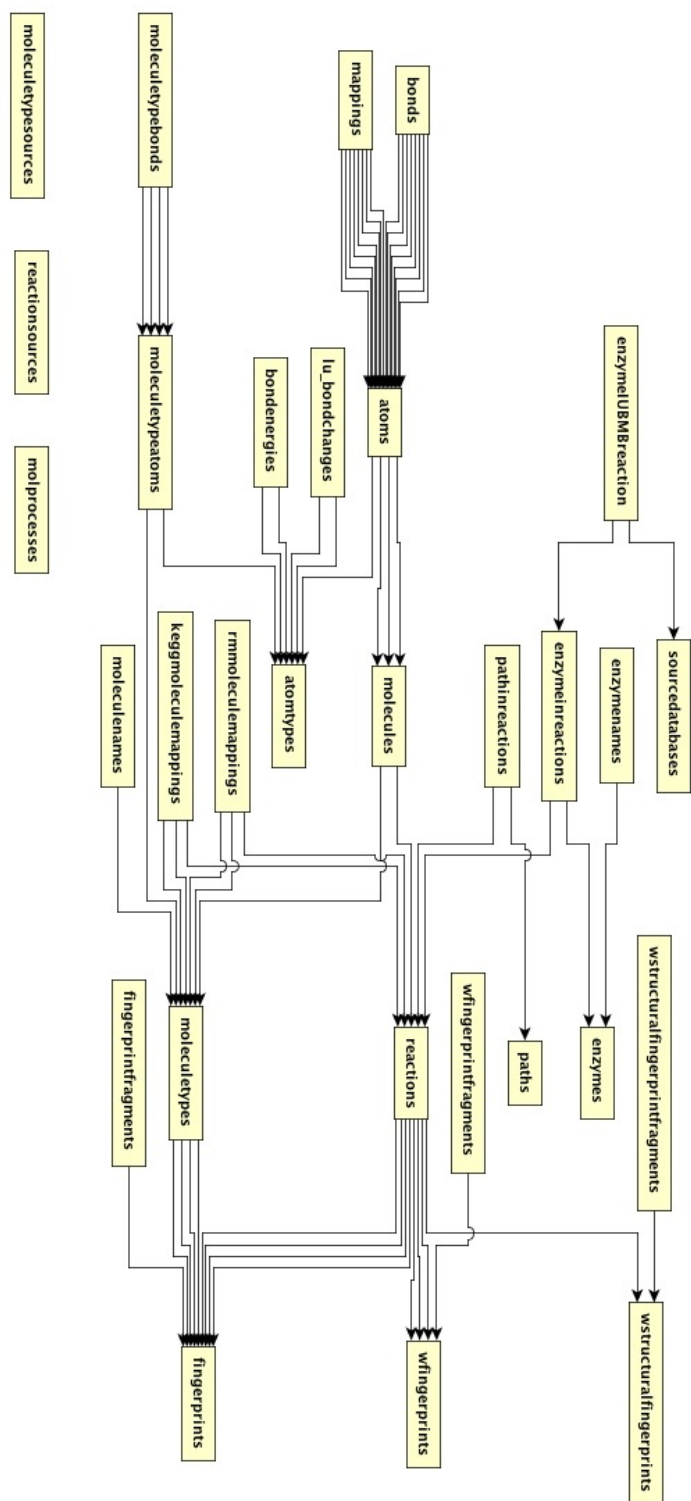


Figure 3.4: Entity Relationship Model

is a mapping record for each atom of a reaction so it is traceable which educt-atom transforms to which product atom. The database also holds information about the enzyme class of the reactions.

Another important information for the ERIS software are the fingerprints for the molecules and reactions, which are used to conduct the different molecules and reaction searches. See Figure 3.4 for the ER Model.

3.4.1 Database wrapper

The database wrapper consists of the three java classes providing the interface between the front end and the database. It can give an id retrieve chemical structures or send statements to get the similarity scores of molecules in the database when given a query molecule.

3.5 Searches

ERIS provides two algorithms to perform searches. Each of them yields an euclidean distance where zero stands for an exact match.

3.5.1 Fingerprints

Fingerprints are used to compare chemical structures. It is a very fast way to do a similarity search and are therefore also used to do a prescreening on the database given a query structure. In ERIS three different types of fingerprints are used.

Molecule fingerprints

In chemoinformatics a fingerprint is considered as a bitset that represents a molecule. Roughly speaking each bit in the fingerprint refers to a specific substructure or another chemical property. A true or false value at his position indicates if this property is present or not. Fingerprints are used to make molecules comparable since it is quite difficult to compare the molecules themselves. Due to their composition it also becomes possible to perform structure screening.

Structure based reaction fingerprints

The structured based reaction fingerprints bitwise sum up the fingerprints of its compounds yielding an array of integers which then can be used to compare reactions in a fast manner.

Bond-change based reaction fingerprints

In order to store the bond changes of a reaction in fingerprints three integer arrays are used.

- CF Fingerprint
bonds cleaved or formed in a reaction
- OC Fingerprint
bonds undergoing an order-change
- ST Fingerprint
undergoing a stereo change

Each position in a fingerprint represents a bond type (for example C-O) and the integer at this position for the count of this type of bond change in the reaction i.e. a 5 at the C-O position of the CF fingerprint would mean that there are 5 C-O bonds formed/cleaved.

3.5.2 Maximum Common Subgraph

An alternative for performing a similarity search is a MCS search where the most common subgraph of a query molecule is searched for. This search is much more computationally intensive and therefore is only used for a smaller amount of target molecules. For example it is used after a prescreening of the database using fingerprints with a certain cutoff score.

3.5.3 Physicochemical Properties

Another searching option that was implemented is the one for physico-chemical properties. The database stores these properties and therefore database wrapper can retrieve the appropriate molecules. It is possible to search for following attributes:

- Molecular weight
- Hydrogen bond acceptor (HBA)
- X log P (Partition coefficient)
- Rotatable bonds
- Atoms with hydrogen

- Hydrogen bond donor (HBD)
- $X \log D$ (Distribution coefficient)
- Count of rings
- Molar refractivity
- Atoms without hydrogen
- Eccentric connectivity
- Topological polar surface area
- Fused rings
- Total charge

Also standard filters like for example Lipinski's Rule of Five can be applied and therefore yield molecules that are likely to be pharmacologically active.

Chapter 4

Results

Since my major part was the development of a front end there are now some screenshots of ERIS.

Figure 4.1 shows a reaction of the alcohol dehydrogenase enzyme. The numbers for each atom map the educt atoms to the product atoms it is possible to see what happens to each atom. The straight lines crossing bonds symbolize a bond broken while the arcs represents a formed bond or in that case an increase of the bond order.

In Figure 4.2 the Reaction matrix of this reactions is shown. As already described in section 3.3 a positive value stands for a formed bond or a bond order increase while a negative value represents a broken bond or a bond order decrease. In the main diagonal the change of the free valence electrons count can be found.

The transition state is shown in Figure 4.3. Here all the bond changes of the reaction can be observed in one single "transition state molecule".

In Figure 4.4 the query molecule of a MCS search can be seen while Figure 4.5 the results in four target molecules can be found.

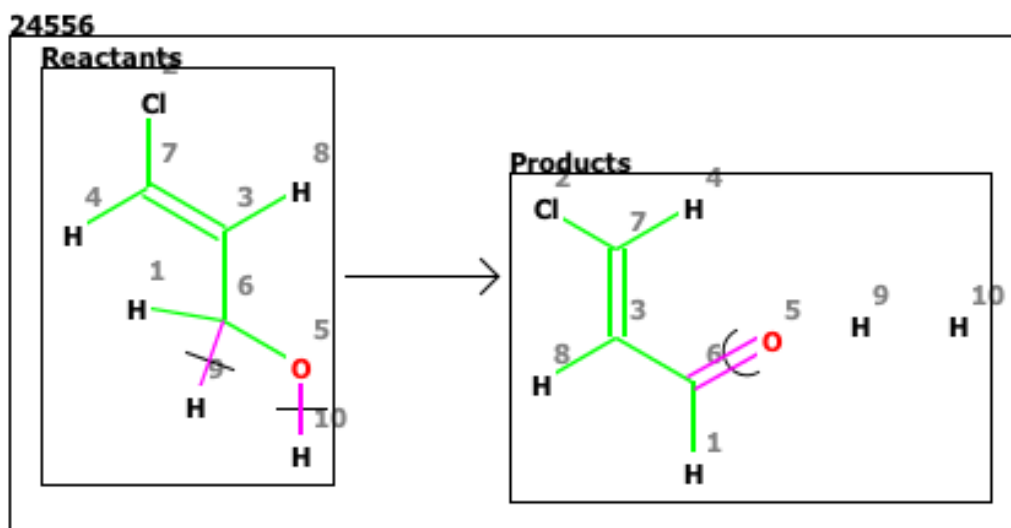


Figure 4.1: Atom Atom Mapping and bond changes of a alcohol dehydrogenase reaction

	H1	C6	C3	O5	H9	C7	H8	H10	Cl2	H4
H1	0	0	0	0	0	0	0	0	0	0
C6	0	0	0	1	-1	0	0	0	0	0
C3	0	0	0	0	0	0	0	0	0	0
O5	0	1	0	0	0	0	0	-1	0	0
H9	0	-1	0	0	1	0	0	0	0	0
C7	0	0	0	0	0	0	0	0	0	0
H8	0	0	0	0	0	0	0	0	0	0
H10	0	0	0	-1	0	0	0	1	0	0
Cl2	0	0	0	0	0	0	0	0	0	0
H4	0	0	0	0	0	0	0	0	0	0

Figure 4.2: Reaction matrix of a alcohol dehydrogenase reaction

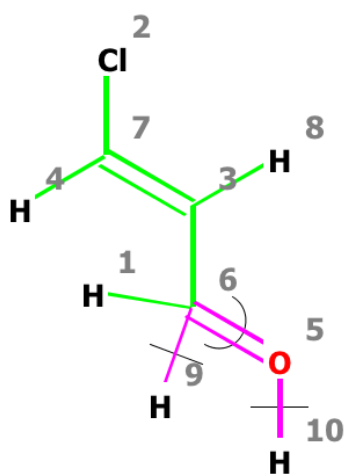


Figure 4.3: Transition state of a alcohol dehydrogenase reaction

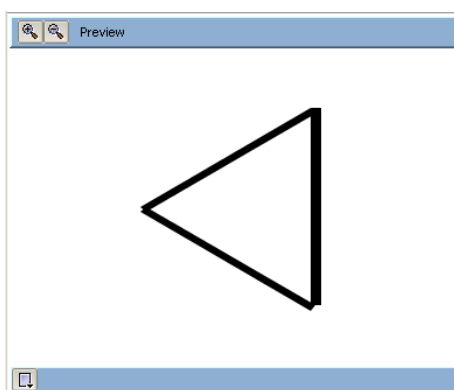


Figure 4.4: MCS Search: Query structure

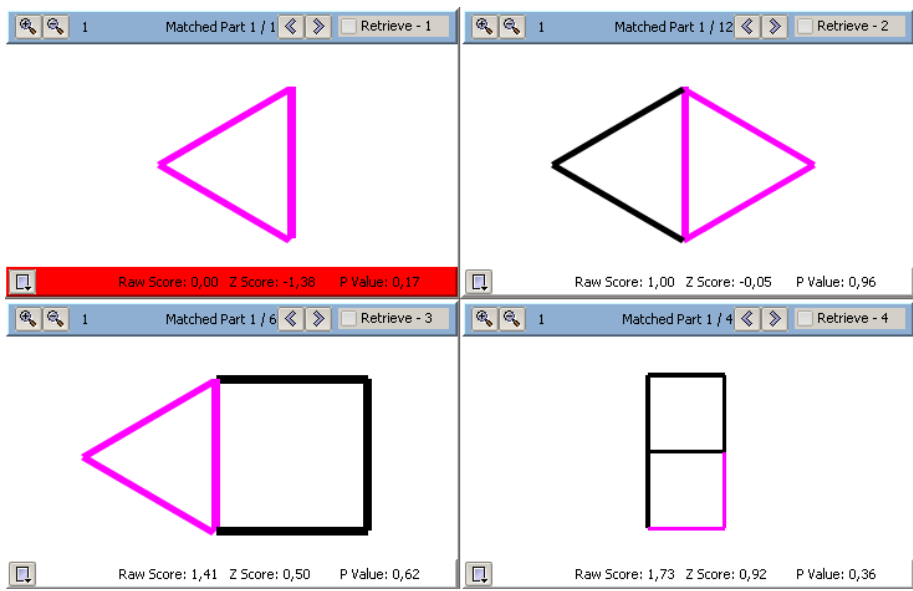


Figure 4.5: MCS Search: Results

Chapter 5

Discussion

5.1 Future improvements

Due to the amount of information stored in the database and the possibilities CDK offers in terms of chemical algorithms there is a huge extension potential for ERIS. In the near future a function to work with reaction chains. It will be possible to map not only educt- to product atom but also product atoms of one reaction to the reactant atoms of its successor. That will give the possibility to trace chemical structure through the whole reaction chain.

Furthermore there will be a possibility to retrieve bar and pie charts based on the type and count of bond changes in a reaction/enzyme class/whole database.

It should also become possible to analyse metabolic pathways and - networks eventually leading to an information system about whole biological networks.

Chapter 6

Acknowledgements

I would like to thank Dr. Syed Asad Rahman, my supervisor at the EBI for offering me a position in the ERIS development team and therefore giving me the opportunity to work in a very interesting field and for his support and patience throughout the whole development.

I am also grateful to Dr. Lorenzo Baldacci for his help and advices and Prof. Janet M. Thornton for working in her group.

Special thanks go to the CDK Development Team and of course to my supervisor at Graz University of Technology Univ.-Prof. Dipl.-Ing. Dr.techn. Zlatko Trajanoski.

Bibliography

- [1] James Dugundji and Ivar Ugi. *Computers in Chemistry*, volume 39/1 of *Topics in Current Chemistry*, chapter An algebraic model of constitutional chemistry as a basis for chemical computer programs, pages 19–64. Springer Berlin / Heidelberg, 1973.
- [2] Christoph Steinbeck, Yongquan Han, Stefan Kuhn, Oliver Horlacher, Edgar Luttmann, and Egon Willighagen. The chemistry development kit (cdk): An open-source java library for chemo- and bioinformatics. *Journal of Chemical Information and Computer Sciences*, 43(2):493–500, 2003.