

EXPERIMENTAL ENZYME DATA AS PRESENTED IN BRENDA - A DATABASE FOR METABOLIC RESEARCH, ENZYME TECHNOLOGY AND SYSTEMS BIOLOGY

IDA SCHOMBURG, ANTJE CHANG, CHRISTIAN EBELING, GREGOR HUHN,
OLIVER HOFMANN, DIETMAR SCHOMBURG*

CUBIC (Cologne University Bioinformatics Centre), Institute of Biochemistry,
Köln, Germany

E-Mail: *d.schomburg@uni-koeln.de

Received: 15th April 2004 / Published 1st October 2004

ABSTRACT

BRENDA represents the most comprehensive information system on enzyme and metabolic information, based on primary literature. The database contains data from at least 83,000 different enzymes from 9800 different organisms, classified in approximately 4200 EC numbers. BRENDA includes biochemical and molecular information on classification and nomenclature, reaction and specificity, functional parameters, occurrence, enzyme structure, application, engineering, stability, disease, isolation, and preparation, links, and literature references. The data are extracted and evaluated from approximately 46,000 references, which are linked to PubMed as long as the reference is cited in PubMed. In the last year BRENDA underwent major changes including a large increase in updating speed with more than 50% of all data updated in 2002 or in the first half of 2003, the development of a new EC-tree browser, a taxonomy-tree browser, a chemical substructure search engine for ligand structure, the development of controlled vocabulary and an ontology for some information fields, and a thesaurus for ligand names. The database is accessible free of charge for the academic community at <http://www.brenda.uni-koeln.de>.

Analysis of the experimental data stored in BRENDA shows a number of problems that prohibit a systematic comparison and evaluation of experimental protein data. This is caused by the fact that on the one hand, many experimental data are determined in a non-systematic way and that - on the other hand - the existing recommendations on nomenclature are systematically ignored by most authors of biochemical and molecular-biological papers. Examples will be given.

INTRODUCTION

Enzymes represent the largest and most diverse group of all proteins, catalysing all chemical reactions in the metabolism of all organisms. They play a key role in the regulation of metabolic steps within the cell. With the recent development and progress of projects of structural and functional genomics and metabolomics, the systematic collection, accessibility and processing of enzyme data becomes even more important in order to analyse and understand biological processes.

The protein function database BRENDA [1] was founded in 1987 at the German National Research Centre for Biotechnology (GBF) and is continued at the Cologne University Bioinformatics Centre (CUBIC). Firstly, BRENDA was published as a series of books (Handbook of Enzymes, Springer [2]). The second edition was started in 2001. Eighteen volumes have been published so far, each containing about 500-600 pages encompassing 50-150 EC classes. By 2006, 15 more volumes will have been produced.

BRENDA contains a very large amount of enzymatic and metabolic data and is updated and evaluated by extracting information from the primary literature. BRENDA represents a comprehensive relational database containing all enzymes classified according to the EC system of the Enzyme Nomenclature Committee (IUBMB [3]). This classification is based on the type of reaction (e.g. oxidation, reduction, hydrolysis, group transfer) catalysed by the enzyme.

All data in BRENDA have a standard structure:

Value (or range of values), e.g. Turnover number

Protein (information on the exact protein, either organism or - if available - sequence)

Literature reference

Commentary (giving experimental conditions, isoform, etc.)

Additional information (e.g. Substrate for kinetic constants, reversibility and product for substrate fields, etc.)

Since 1998 all data are available on the internet in a relational database system. Since then the user interface has been developed intensively to provide a sophisticated access to the data. The user can choose from seven search modes:

Experimental Enzyme Data as Presented in BRENDA

Quick search, Full text search, Advanced search, Substructure search, TaxTree search, ECTree browser, and Sequence search. In 2003 a BRENDA discussion forum was started.

Access to BRENDA is free for the academic community at <http://www.brenda.uni-koeln.de>. An in-house version for academic users is available for a low handling fee. Commercial users are required to purchase a license.

PHILOSOPHY

In contrast to other databases, BRENDA is not limited to a specific aspect of the enzyme or to a specific organism. It covers organism-specific information on functional and molecular properties, enzyme names, catalysed reaction, occurrence, sequence, kinetics, substrates/products, inhibitors, cofactors, activators, structure and stability. Presently, BRENDA holds information on 4200 enzyme classes, which represent more than 83,000 different enzyme molecules. Since 2002 the annotation speed has been tripled to 1000 enzyme classes per year.

THE ANNOTATION PROCEDURE

The annotation procedure comprises the insertion of new data, the reallocation and reclassification of enzymes to their respective class and the removal of data which have been proven to be wrong. The data are annotated manually and are controlled for consistency via computer-aided and manual techniques. Special sections of BRENDA contain automatically annotated data which are indicated explicitly.

Step 1 Enzyme Names

The first step is the search for all the names with which the enzyme is associated. Enzyme names can be found at the IUBMB, from databases, or from the literature.

Step 2 Literature evaluation

In the literature evaluation step the major databases (CAS [4], PubMed [5], databases for specific protein classes) are searched for literature dealing with the respective enzyme class. The number of citations varies greatly with the enzyme class, some searches will result in only a single publication for an enzyme class, others may produce 10,000 hits. In a series of refinement cycles these search results are reduced to those references which will probably yield information that is suitable for at least one of BRENDA's information fields. Manual assessment of the title or the abstract is often necessary to make the right choice.

Step 3 Annotation and quality control

The annotation involves a high amount of manual work because the literature references rarely contain all relevant data in concise tables. Another great amount of work is required to sort out all the various names for enzymes and chemical compounds. Quite often in this stage of annotation, the literature reveals data for enzymes which are not yet classified in the EC number system. These are then collected, completed via an exhaustive search and assembled as a proposal for a new entry in the list of enzymes at the IUBMB. After approval a new EC number is awarded which is then integrated into BRENDA.

The revision and annotation process frequently reveals inconsistencies regarding the enzyme's classification. Then the respective enzyme will be allocated to a different enzyme class after the IUBMB has given approval.

AUTOMATIC CONTROL OF DATA (selected)

- no data-fields missing?
- EC-Number correct?
- all references cited?
- all organisms cited?
- entries in numeric fields in the correct range?
- all brackets, braces, parentheses correct?
- structure of commentary correct?
- journal abbreviations according to list?
- all organisms cited with their correct references and vice versa?
- names for organisms in accordance to *NCBI taxonomy*?
- CAS Number correct?
- correct terms in fields application, post-translational modification?

In addition for a number of fields a controlled vocabulary was introduced and is checked during processing time (application, cofactors, localization, organic solvent stability, post-translational modification, reaction type, source tissue, subunits).

Step 4 Processing the database

In consecutive final steps the data are processed for integration into the database.

Compilation of BRENDA database:

- Parsing of TEXT data, integration into non-organism-specific database, final automatic control
- Split up of database into multiple tables with organism-specific information.

Compilation of BRENDA LIGAND database:

- draw structures of new ligands (Mol-format)
- convert to SMILES
- create thesaurus
- convert mol-files to gif-images.

THE BRENDA DATA STRUCTURE

- Classification and Nomenclature
- Reaction & Specificity
- Functional Parameters
- Organism related Information
- Enzyme Structure
- Isolation and Preparation
- Literature References
- Application and Engineering
- Enzyme-Disease Relationship

CLASSIFICATION AND NOMENCLATURE

Since enzyme names have a long history they are not unique. In many cases the same enzymes became known by several different names, while conversely the same name was sometimes given to different enzymes. Many of the names conveyed little or no idea of the nature of the reactions catalysed, and similar names were sometimes given to enzymes of quite different types.

The International Commission on Enzymes was founded in 1956 by the International Union of Biochemistry. Since then the system of EC numbers with systematic names and recommended names has been established.

Currently there are 3741 active EC numbers plus 556 numbers for deleted or transferred enzymes. The old numbers have not been allotted to new enzymes; instead the place has been left vacant or comments are given concerning the fate of the enzyme (deletion or transfer).

In the EC number system an enzyme is not defined by its name but by the reaction it catalyses. In some cases where this is not sufficient, additional criteria are employed such as cofactor specificity or stereospecificity of the reaction. The 3741 active EC numbers currently account for 28,900 synonyms.

THE ENZYME NOMENCLATURE PROBLEM

Unlike other protein classes, a standard nomenclature and recommended names exist for enzymes. Unfortunately they are often not used by researchers in publications. Therefore, often many different names are in use for enzymes, EC 3.1.21.4, i.e. "type II site specific deoxyribonuclease" with 370 different names. Thus, if a researcher searches in literature databases (e.g. PubMed) only those references will be found which are stored with the synonym he uses. The particular name chosen may be in fact a rarely used synonym and thus he will retrieve only a fraction of the information. Table 1 contains examples of enzymes which are characterized by manifold synonyms.

One important aspect of BRENDA data input is to give the user complete information for an enzyme when he queries the database with a single synonym. Thus great effort is invested in the best possible completeness of enzyme names. The majority of the names are extracted manually from the original literature and completed by searching internet databases (e.g. CAS, PubMed, SwissProt).

Table 1. Enzymes with manifold names in BRENDA.

EC-Number	Recommended Name	Number of Synonyms
3.1.21.4	type II site-specific deoxyribonuclease	369
3.1.3.48	protein-tyrosine-phosphatase	169
1.6.5.3	NADH dehydrogenase (ubiquinone)	162
2.7.7.6	DNA-directed RNA polymerase	91
3.1.2.15	ubiquitin thiolesterase	81

Experimental Enzyme Data as Presented in BRENDA

2.7.1.69	protein-Npi-phosphohistidine-sugar phosphotransferase	80
5.2.1.8	peptidylprolyl isomerase	111
3.1.3.16	phosphoprotein phosphatase	74
3.2.1.4	cellulase	72
3.1.1.1	carboxylesterase	60
3.6.3.14	methylphosphotrioglycerate phosphatase	56

THE UNIQUENESS PROBLEM

Another problem in enzyme literature is that often identical names or abbreviations are applied for more than one enzyme thus creating confusion. Moreover the use of ambiguous names would create completely misleading results. In many cases names or abbreviations refer to more than one EC number (Figs 1-3), e.g. The name GTPase applies to 6 different EC numbers within the same subclass, the abbreviation FDH applies to 8 EC numbers in 3 different subclasses, or the name chondroitinase applies to 5 different EC numbers in 2 different EC classes. Thus the use of any arbitrary enzyme name can lead to great confusion and misleading results in the selection of enzyme data from a database.







EC Number	Recommended Name	Synonyms
 3.6.1.46	heterotrimeric G-protein GTPase	GTPase
 3.6.1.47	small monomeric GTPase	GTPase
 3.6.1.48	protein-synthesizing GTPase	GTPase
 3.6.1.49	signal-recognition-particle GTPase	GTPase
 3.6.1.50	dynamine GTPase	GTPase
 3.6.1.51	tubulin GTPase	GTPase

Figure 1. Enzyme classes carrying the name GTPase.









EC Number	Recommended Name	Synonyms
 1.1.1.1	alcohol dehydrogenase	FDH
 1.1.1.122	D-threo-aldose 1-dehydrogenase	FDH
 1.1.99.11	fructose 5-dehydrogenase	FDH
 1.2.1.1	formaldehyde dehydrogenase (glutathione)	FDH
 1.2.1.2	formate dehydrogenase	FDH
 1.2.1.43	formate dehydrogenase (NADP)	FDH
 1.2.1.46	formaldehyde dehydrogenase	FDH
 1.5.1.6	formyltetrahydrofolate dehydrogenase	FDH

Figure 2. Enzyme classes carrying the name FDH.






EC Number	Recommended Name	Synonyms
 3.1.6.4	N-acetylgalactosamine-6-sulfatase	chondroitinase
 3.1.6.12	N-acetylgalactosamine-4-sulfatase	chondroitinase
 3.2.1.35	hyaluronoglucosaminidase	chondroitinase
 4.2.2.4	chondroitin ABC lyase	chondroitinase
 4.2.2.5	chondroitin AC lyase	chondroitinase

Figure 3. Enzyme classes carrying the name chondroitinase.

In BRENDA the information on enzyme nomenclature can be retrieved from the section Classification & Nomenclature which is divided into the data-fields:

- Enzyme Names 34,509 entries
- EC Number 4293 entries
- Recommended/Common Names 4293 entries
- Systematic Names 3425 entries
- Synonyms 27,903 entries
- CAS Registry Number 3955 entries

Any query in BRENDA will have the EC number as the result, thus enabling the user to select the correct enzyme.

REACTION AND SPECIFICITY - METABOLITES AND LIGANDS

An enzyme is defined by the reaction it catalyses. Thus all proteins found to catalyse a specific reaction are summarized under one EC number. Apart from this an enzyme may have a wider substrate specificity and may accept different substrates. These appear in BRENDA in the section Substrate/Product and Natural Substrate/Natural Product.

Additional sections provide lists of inhibitors, cofactors and activating compounds. Since in biological sciences very often trivial names are used instead of IUPAC nomenclature many compounds are known by a variety of names. Thus even simple molecules may have a dozen or more names. For example, the inhibitor 2,2'-bipyridine is cited with 12 different names. BRENDA is equipped with a thesaurus for ligand names. This thesaurus is based on the generation of unique and chiral SMILES-strings [6, 7] for ligand structures in the database.

If the function of a compound is not known, it can be searched in the table LIGANDS. This will perform a search in all data-fields which contain ligand names (substrates, products, natural substrates, inhibitors, cofactors, activating compounds, K_M , K_i).

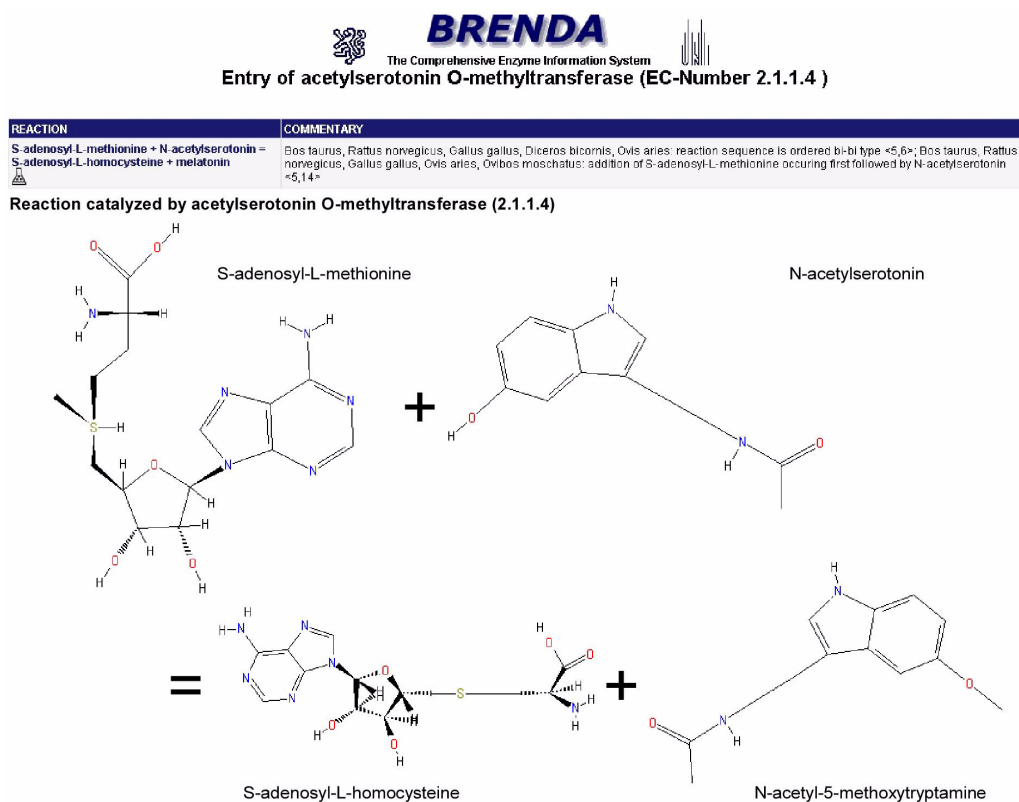


Figure 4. Display of enzyme-catalysed reactions in BRENDA.

The most exhaustive search for ligands is a full-text search of the complete database. This mode, however, does not apply the thesaurus for molecule names. Ligands can also be viewed as 2D-structures thus offering an unambiguous method to display a reaction. (Fig. 4).

Overview BRENDA ligand data

• enzyme/ligand relationships	537,293
• Cofactor	9572
• Activating Substance	10,600
• Metals/Ions	15,641
• Substrates	255,270
• Products	237,686
• Natural Substrate	16,148
• Inhibitors	72,982
• different ligand names	54,895
• ca. 5000 of these macromolecules, molecule classes etc.	
• ligand structures as mol-files	36,820
• Ligand name thesaurus	
• Grouped into 25198 different compounds	

ENZYME FUNCTION AND STABILITY

Functional Parameters

The BRENDA database contains a section for functional parameters of enzymes with these datafields:

• Functional Parameters	116,130
• K_M -value	47,299
• Turnover number	8010
• K_i -value	4441
• Temperature optimum	7762
• Temperature range	1826
• pH optimum	18,086
• pH range	4825
• p_I -value (coming soon)	

Experimental Enzyme Data as Presented in BRENDA

Each of these data-fields is divided into subsections.

Example K_M -Value

- Value
- Substrate
- Organism
- Protein (Swissprot/Trembl Code if available)
- Commentary
 - Experimental conditions
 - Isoform
 - Method
 - Other commentaries
- Literature reference
- Date of last change

An example of entries for turnover numbers can be seen in Figure 5.



Entry of alcohol dehydrogenase (EC-Number 1.1.1.1)


















TURNOVER NUMBER	TURNOVER NUMBER MAXIMUM	SUBSTRATE	ORGANISM	COMMENTARY	LITERATURE	IMAGE
18480	-	ethanol	Equus caballus	wild-type enzyme Adh 1 <92>	92	 2D-image
12800	-	2-buten-1-ol	Rattus norvegicus	isoenzyme ADH-1, pH 10.0 <49>	49	 2D-image
7980	-	cinnamyl alcohol	Equus caballus	wild-type enzyme Adh 1 <92>	92	 2D-image
7380	-	ethanol	Equus caballus	wild-type enzyme Adh 1 <92>	92	 2D-image
7080	-	butanol	Equus caballus	wild-type enzyme Adh 1 <92>	92	 2D-image
6120	-	ethanol	Equus caballus	mutant enzyme W54L <92>	92	 2D-image
5880	-	cinnamyl alcohol	Equus caballus	mutant enzyme W54L <92>	92	 2D-image
5380	-	benzyl alcohol	Rattus norvegicus	isoenzyme ADH-1, pH 10.0 <49>	49	 2D-image
3760	-	ethanol	Rattus norvegicus	isoenzyme ADH-1, pH 10.0 <49>	49	 2D-image
3650	-	1-octanol	Rattus norvegicus	isoenzyme ADH-1, pH 10.0 <49>	49	 2D-image
3396	-	hexaldehyde	Equus caballus	-	42	 2D-image
3372	-	pentanol	Equus caballus	wild-type enzyme Adh 1 <92>	92	 2D-image
3186	-	propan-2-ol	Equus caballus	wild-type enzyme Adh 1 <92>	92	 2D-image
2930	-	1-butanol	Rattus norvegicus	isoenzyme ADH-1, pH 10.0 <49>	49	 2D-image
2930	-	1-pentanol	Rattus norvegicus	isoenzyme ADH-1, pH 10.0 <49>	49	 2D-image
2514	-	propionaldehyde	Equus caballus	-	42	 2D-image
2124	-	butyraldehyde	Equus caballus	-	42	 2D-image
2090	-	butanol	Homo sapiens	-	53	 2D-image
1962	-	butanol	Equus caballus	mutant enzyme W54L <92>	92	 2D-image
1908	-	acetaldehyde	Equus caballus	-	42	 2D-image
1840	-	ethanol	Homo sapiens	-	53	 2D-image
1776	-	hexanol	Equus caballus	wild-type enzyme Adh 1 <92>	92	 2D-image
1770	-	benzaldehyde	Equus caballus	-	42	 2D-image

Figure 5. Sample of turnover numbers in BRENDA.

The data are often obtained under very different experimental conditions. Since every laboratory carries out experiments on enzyme characterizations under individually defined conditions, and since they depend on the given experimental know-how, methods and technical equipment available, raw data for the same enzyme from different laboratories are not at all comparable. Therefore BRENDA not only contains individual values but very often the experimental conditions are also included. Because until now there has been no standardization for documenting these, the details are only given as text. Each entry is linked to a literature reference, thus for reproduction of data the researcher may have to go back to the original literature.

Example: K_M and pH optimum for the human enzymes of glyceraldehyde-3-phosphate dehydrogenase (phosphorylating) (EC-Number 1.2.1.12).

K_M

- 0.002** 3-phospho-D-glyceroyl phosphate, enzyme form E6.8, pH 7 <15>
- 0.03** 3-phospho-D-glyceroyl phosphate, enzyme form E9.0, pH 9 <15>
- 0.14** 3-phospho-D-glyceroyl phosphate, <2>
- 0.17** 3-phospho-D-glyceroyl phosphate, enzyme form E6.8, pH 9 <15>

pH optimum

- 7** enzyme form E6.8, two pH optima: pH 7.0 and pH 8.5, with activity between pH 7.5 and pH 8.0 being rather low <15>
- 7.2-7.3** reaction with 3-phospho-D-glyceroylphosphate <34>
- 8.5** enzyme form E6.8, two pH optima: pH 7.0 and pH 8.5, with activity between pH 7.5 and pH 8.0 being rather low <15>
- 8-8.3** reaction with D-glyceraldehyde 3-phosphate <34>
- 9.8** enzyme form E8.5, D-glyceraldehyde 3-phosphate <15>

 Experimental Enzyme Data as Presented in BRENDA

These data do not allow automatized access and are unsuitable for the modelling of sections of the cellular metabolism, the whole cellular metabolism or the interaction of cells within tissues and organs. Thus a new data model is needed which provides data which have been generated under standardized experimental conditions.

Stability Parameters

In BRENDA the stability of the enzymes is documented in six sections

• Stability parameters	27,154
• pH stability	3755
• Temperature stability	8841
• General stability	5702
• Organic solvent stability	452
• Oxidation stability	452
• Storage stability	7951

Stability data are especially difficult to put into an automatically interpretable format since the literature data are very inhomogeneous. Whereas one research group states an enzyme to be stable at a certain temperature another will find it to be highly unstable. The discrepancy may be due to the type of buffer, presence of substrates, cofactors, stabilizing or destabilizing ingredients, type of storage vial.

Even some of the purification steps can result in a lower or higher stability. Therefore the stability data in BRENDA are as detailed as possible, reproducing details from the literature.

The sections on General stability and Storage stability contain the organism, text describing conditions, a time and a reference. These two sections contain rather inhomogeneous information for which a standard format has not yet been found.

The sections on pH stability and Temperature stability contain a value, the organism, a commentary and a literature reference. Looking at the value alone will not give sufficient information because the enzyme may have varying stabilities depending on the presence of buffer components or stabilizing/destabilizing agents.

Standardization of experimental conditions

Standardization of experimental conditions is a prerequisite for two reasons:

1. In order to render kinetic or stability data comparable they must be obtained under identical experimental conditions. It is impossible to compare the efficiency of two enzymes if their reaction has been monitored at different pH values which may not even be the optima. Also the stability of enzymes can only be compared if the conditions are identical. In BRENDA ca. 50% of the K_M values are measured at physiological pH values, ca. 33% refer to natural substrates.
2. For the creation of metabolic networks the kinetic data must represent the enzyme's reaction under physiological conditions. These have to be defined regarding the temperature, the pH, ionic strength, or macromolecular crowding. Assay procedures and assay conditions need to be the same to obtain comparable data.

Organism-related information

For the organisms in BRENDA the taxonomy-lineage is given if the respective organism can be found in the NCBI taxonomy database. Using the TaxTree search mode the user can search for enzymes along the taxonomic tree and move to higher or lower branches to either get an overview or to restrict the search.

The tissue may be an important criterion for an enzyme. Sometimes enzymes are restricted to a single tissue or a tissue may express a tissue-specific isoenzyme. The BRENDA tissues are grouped into a hierarchical ontology which was developed especially for this database.

The localization terms are in accordance with the terms of the Gene Ontology [9] consortium.

Overview organism-related data:

• Organism/enzyme relationships	69,408
• from 6728 different organisms	
• Source Tissue/enzyme relationships	25,482
• for 1408 different tissues and cell-lines	
• Localization/enzyme relationships	10,973
• for 148 different subcellular locations	

ENZYME STRUCTURE

Whereas the SwissProt and PDB links are automatically generated, the sections molecular weight, subunits and post-translational modifications are extracted manually from the literature. As the accuracy of the value for the molecular weight or the size of the subunits is dependent on the method of determination, BRENDA gives the method in the commentary, if available. Of the 3741 EC classes sequences are only available for 2166 classes.

Overview enzyme structure

• SwissProt links	53,999
• PDB links	10,610
• Molecular weight	17,715
• Subunit	10,744
• Posttranslational modification	19,019

Isolation and Preparation

The isolation/purification section contains information on purification, crystallization, cloning and renaturation. Due to the inhomogeneity of the data, these are in non-structured text-format.

Overview isolation and preparation

• Purification	14,380
• Cloned	5491
• Renaturation	317
• Crystallization	1548

LITERATURE REFERENCES

For the BRENDA database all information except the sequence information and the enzyme-associated diseases is manually extracted from 50,300 scientific publications. The major drawback of this method is the low speed of annotation compared to automated methods. During the manual annotation procedure the scientist is able to assess the facts, compare the results of different research groups, and choose the data which he wants to include in BRENDA depending on the experimental conditions. For example data obtained with a crude cell extract have to be distinguished from data that were obtained with a purified enzyme.

Studying the literature of an enzyme sometimes reveals misclassification and thus leads to the transfer of an enzyme to another enzyme class.

METABOLIC DISORDER-RELATED INFORMATION

BRENDA contains a large section of data for metabolic disorders which are connected to a dysfunction of an enzyme. However, due to the rapid growth of information there is a widening gap between manually annotated data and information available in the literature. In order to alleviate the problem a tool to automatically extract enzyme-related information from the biomedical literature was developed. It is based on the co-occurrence of enzyme names and interesting phrases which are identified utilizing concepts from the Unified Medical Language System (UMLS) [12]. A variety of filters reduce the number of false extraction events, among them a classification of sentences based on their semantic context by a Support Vector Machine (KMO) [13].

A prototype of this concept based approach links 524 enzyme classes from the BRENDA database to more than 1400 disease related concepts, achieving a precision of more than 90% and a recall of 49% on a test-set of 1500 manually annotated sentences. Current work is focusing on expanding the scope of the tool to include other fields of interest, i.e. subcellular localization of enzymes or co-occurrences of enzyme names with pharmaceutical compounds.

Overview Disease-related data

- ca. 50,000 PubMed references with disease-term and enzyme name in title
- ca. 20,000 references selected by text-mining tool
- 506 EC numbers in disease-related papers
- 1407 disease terms related to enzymes

APPLICATION AND ENGINEERING

- | | |
|---------------|------|
| • Application | 1413 |
| • Engineering | 4531 |

Enzymes are widely applied in industry, pharmacology, medicine or for analytical purposes. BRENDA not only lists established applications but also putative future usages.

This data field is based on a controlled vocabulary, the comments are in text format. The engineering section displays the amino acid exchange in the engineered enzyme. The comments give as much detail on the properties of the mutated enzyme as available, which is mostly restricted to a short comment on the activity or stability. For mutants with kinetic constants these can be found in the functional parameters section.

SUMMARY AND PERSPECTIVES

The enzyme database BRENDA represents data for ~4000 enzyme classes defined in the EC system. The data give detailed information on nomenclature, specificity, structure, organism, functional parameters, enzyme stability and diseases related to dysfunction. All data are linked to primary literature references. Enzyme data are essential for understanding and predicting the biological chemistry of the cell. For a reliable interpretation of these values by computational methods standardization is indispensable:

1. All enzymes names must be in accordance to the IUBMB system of enzyme nomenclature.
2. Thermodynamic and kinetic data must be recorded under defined conditions, mimicking physiological conditions.
3. Metabolites must carry unequivocal names or identifiers.
4. Organisms and cell-types, tissues and cellular components must be named in accordance to defined ontologies.

REFERENCES

- [1] Schomburg, I., Chang, A., Ebeling, E., Gremse, M., Heldt, C., Huhn, G., Schomburg, D. (2004) BRENDA, the enzyme database: updates and major new developments. *Nucl. Acids Res.* **32** : D431-D433.
 - [2] Schomburg, D., Schomburg, I. (2001) *Springer Handbook of Enzymes*, 2nd Edn. Springer, Heidelberg, Germany.
 - [3] *Enzyme Nomenclature* (1992) Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes, NC-IUBMB. Academic Press, New York.
 - [4] Ridley, D.D. (2002) *SciFinder and SciFinder Scholar*. J. Wiley & Sons, New York
-

-
- [5] Wheeler, D.L., Church, D.M., Lash, A.E., Leipe, D.D., Madden, T.L., Pontius, J.U., Schuler, G.D., Schriml, L.M., Tatusova, T.A., Wagner, L. Rapp, B.A. (2001) Database resources of the National Center for Biotechnology Information. *Nucl. Acids Res.* **29**: 11-16.
- [6] Weininger, D. (1988) SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **28**: 31-36.
- [7] Weininger, D., Weininger, A., Weininger, J. (1989) SMILES. 2. Algorithm for generation of unique SMILES notation. *J. Chem. Inf. Comput. Sci.* **29**: 97-101.
- [8] Steinbeck, C., Han, Y., Kuhn, S., Horlacher, O., Luttmann, E., Willighagen E. (2003) The Chemistry Development Kit (CDK): An open-source java library for chemo- and bioinformatics. *J. Chem. Inf. Comput. Sci.* **43**(2):493-500.
- [9] Ashburner, C.A., Ball, J.A., Blake, D., Botstein, H., Butler, J.M., Cherry, A.P., Davis, K., Dolinski, S.S., Dwight, J.T., Eppig, M.A. et al. (2000) Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**: 25-29.
- [10] Berman, H.M., Westbrook, J., Feng, Z., Gillilan, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E. (2000) The Protein Data Bank. *Nucl. Acids Res.* **28**: 235-242.
- [11] Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S., Schneider, M. (2003) The Swiss-Prot protein knowledgebase and its supplement TrEMBL in 2003. *Nucl. Acids Res.* **31**: 365-370.
- [12] Bodenreider, O. (2003) The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucl. Acids Res.* **32**(database issue): 267-270.
- [13] Kazama, J., Makino, T., Ohta, F., Tsujii, J. (2002) Tuning support vector machines for biomedical named entity recognition. *Proceedings of the workshop on natural language processing in the biomedical domain*, Philadelphia, pp.1-8.
-