BIOINFORMATICS APPLICATIONS NOTE Vol. 23 no. 15 2007, pages 2015–2017
doi:10.1093/bioinformatics/btm268

Sequence analysis

The Cytochrome P450 Engineering Database: a navigation and prediction tool for the cytochrome P450 protein family

Markus Fischer[1], Michael Knoll[2], Demet Sirim[2], Florian Wagner[2], Sonja Funke[2] and Juergen Pleiss[2,*]

[1]Department of Biochemistry & Molecular Biophysics, Columbia University, 1130 St. Nicholas Ave, New York, NY 10032, USA and [2]Institute of Technical Biochemistry, University of Stuttgart, Allmandring 31, 70569 Stuttgart, Germany

Received on March 20, 2007; revised on May 7, 2007; accepted on May 10, 2007

Advance Access publication May 17, 2007

Associate Editor: Alex Bateman

ABSTRACT

Summary: The Cytochrome P450 Engineering Database (CYPED) has been designed to serve as a tool for a comprehensive and systematic comparison of protein sequences and structures within the vast and diverse family of cytochrome P450 monooxygenases (CYPs). The CYPED currently integrates sequence and structure data of 3911 and 25 proteins, respectively. Proteins are grouped into homologous families and superfamilies according to Nelson's classification. Nonclassified CYP sequences are assigned by similarity. Functionally relevant residues are annotated. The web accessible version contains multisequence alignments, phylogenetic trees and HMM profiles. The CYPED is regularly updated and supplies all data for download. Thus, it provides a valuable data source for phylogenetic analysis, investigation of sequence–function relationships and the design of CYPs with improved biochemical properties.

Abbreviations: Cytochrome P450 Engineering Database, CYPED; cytochrome P450 monooxygenase, CYP; Hidden Markov Model, HMM.

Availability: www.cyped.uni-stuttgart.de

Contact: Juergen.Pleiss@itb.uni-stuttgart.de

1 INTRODUCTION

Cytochrome P450 monooxygenases (CYPs) are heme containing enzymes that metabolize physiologically important compounds in many species of microorganisms, plants, animals and humans. CYPs catalyse the oxidation of a wide range of endogenous compounds in biosynthetic and biodegradation pathways, as well as xenobiotics such as drugs and environmental contaminants (Montellano, 1995). Therefore, understanding the substrate specificities of human CYPs is crucial for successful drug development (Raucy *et al.*, 2001). Although there are already numerous resources dedicated to CYPs like the Cytochrome P450 Homepage (http://drnelson.utmem.edu/CytochromeP450.html), the Homepage of the Human Cytochrome P450 (CYP) Allele Nomenclature Committee (http://www.imm.ki.se/CYPalleles/), the Directory of P450-containing Systems (http://www.icgeb.trieste.it/), the P450 Knowledgebase (http://cpd.ibmh.msk.su/) (Lisitsa *et al.*, 2001), the Arabidopsis P450 database (http://www.p450.kvl.dk/), the Insect P450 Site (http://p450.antibes.inra.fr/) or the P450s in PROMISE (http://metallo.scripps.edu/PROMISE/P450.html), a common data structure enabling the integration of available information on protein sequence and structure is still lacking. Therefore the Cytochrome P450 Engineering Database (CYPED) was implemented using the data warehouse system DWARF (Fischer *et al.*, 2006). The underlying data model assists the systematic analysis of the relationship of sequence, structure, and function of this vast and highly diverse protein family. The CYPED is the first cytochrome P450 data resource that combines information on sequences, sequence alignments, annotation and structures of CYPs. For data retrieval sequence, structure and annotation information is extracted from GenBank (Benson *et al.*, 2003) and PDB (http://www.pdb.org/). Functional annotation information is extended by an automated annotation transfer and was manually validated and enriched. Besides, the online accessible version, which is publicly available, supports the classification of unknown sequences by performing a BLAST search against the CYPED or by alignment to family-specific HMM profiles.

2 DEVELOPMENT AND CONSTRUCTION

Seed sequences of CYPs were extracted from the Cytochrome P450 Homepage (Nelson *et al.*, 2002) and assigned to homologous families and superfamilies according to the Nelson classification scheme (Nelson, 2006). For each seed sequence, a BLAST search (Altschul *et al.*, 1997) was performed in the non-redundant sequence database at GenBank (Benson *et al.*, 2003) with a low E-value ($E = 10^{-100}$) to prevent overlapping hits among different superfamilies. For each hit, information on sequence, position-specific annotations, functional descriptions and the source organism was extracted and loaded by an automated retrieval system into an in-house developed relational database system (Fischer and Pleiss, 2003). New protein entries are assigned to homologous families and superfamilies according to their

*To whom correspondence should be addressed.

sequence similarity. The parameters are chosen as specified by Nelson. Proteins sharing a sequence identity of ≥40 or ≥55% are members of the same superfamily or homologous family, respectively. About 2% of sequences were individually assigned to a family according to the recommendations of the P450 Nomenclature Committee. This procedure was applied to sequences which do not share a identity of ≥40% with members of a homologous family but still belong to this family by definition. About 30% of the proteins within the database have not been classified by the nomenclature committee yet, and thus are assigned to the corresponding family by sequence similarity and named as 'homologous protein of family X (by similarity)'. They will be reassigned automatically during a database update in case of the classification information changes. Therefore, in contrast to existing P450 resources, the CYPED includes additional information which is expected to deepen the understanding in sequence–structure–function relationship of the CYP protein family and to apply the new gained knowledge to the design of improved CYPs.

CYP sequences that originate from the same organism and share a sequence identity of at least 98% are assigned to a single protein entry. For each protein entry the longest sequence was defined as reference sequence of the respective protein. For GenBank entries representing protein structures, monomers were extracted from the ExPDB database (Schwede *et al.*, 2000) and deposited as structure entries. Secondary structure information was calculated using DSSP (Kabsch and Sander, 1983), stored and annotated. To improve consistency and quality of the data, the classification into families and superfamilies for those protein entries that have not been classified was validated by performing multisequence alignments and a phylogenetic analysis. Multisequence alignment was also used to enrich annotation information and to control annotation quality. It was assumed that conserved sequence motifs should align for each superfamily. Therefore annotation information was transferred from one sequence to all other sequences in an alignment if the respective residues were conserved.

The CYPED is updated regularly. By an automated Perl script (Fischer *et al.*, 2006) new sequences are retrieved. New annotation information at GenBank for existing sequence entries and structure information is updated as well. Additionally, the update script takes care of changes of classification information.

3 DATA CONTENT AND ANNOTATED MULTISEQUENCE ALIGNMENTS

The CYPED contains sequence data on 3911 proteins. For 25 proteins of 20 different homologous families crystal structures are deposited. Since the CYPED provides a protein analysis tool to investigate the relationship between protein sequences, structures and their function, the data content is limited to sequences, structures and annotation information on amino acid level. The protein entries are assigned to 1111 homologous families, which are grouped into 531 superfamilies. Superfamilies and homologous families are represented as multisequence alignments generated by CLUSTALW (Thompson *et al.*, 1994).

Common CYP motifs not present in GenBank entries were extracted from literature and annotated within CYPED entries. These motifs are the proline-rich region at the N terminus of CYPs (Kemper, 2004), the motif at the C-terminal end of helix K, the AGXXT motif present in helix I, and the functionally essential cysteine (Mestres, 2005). Amino acids linked to functional annotations are coloured within the alignments, and the residue number and further information is displayed upon moving the cursor over the respective amino acid. For each alignment, each column is coloured by the amino acid conservation score as calculated by PLOTCON (Rice *et al.*, 2000). For each homologous family and superfamily family-specific HMM profiles (http://hmmer.janelia.org/) are supplied.

4 WEB ACCESSIBILITY

The CYPED is accessible at http://www.cyped.uni-stuttgart.de by any JavaScript capable WWW browser. It can be browsed by superfamilies and homologous families, organisms, protein structures and the systematic CYP nomenclature. Protein names, source organisms, identifier codes and links to the corresponding GenBank entries are presented as tables. Protein sequences and trees can either be displayed with their accession codes as identifiers or their systematic names and source organisms. The hits are linked to the respective sequence entry, superfamily and homologous family. This functionality can also be applied to classify unknown sequences by performing a BLAST search on the web interface against the CYPED or by alignment to family-specific HMM profiles, which can be downloaded. All multisequence alignments, phylogenetic trees and structural monomers have been pre-calculated, and can be visualized or downloaded from this web site, too. Additionally, an archive can be downloaded comprising sequences, structures, alignments and phylogenetic trees grouped by families, and a formatted text file listing all protein information.

ACKNOWLEDGEMENTS

We acknowledge valuable contributions by Michael Krahn. This work was supported by the German Federal Ministry of Education and Research (project PTJ 0313080) and the German Research Foundation (SFB 706). We further are grateful for the unknown reviewer's comments which contributed greatly to the improvement of data content and the user interface of the CYPED. Funding to pay the Open Access publication charges was provided by the German Research Foundation.

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Benson,D.A. *et al*. (2003) GenBank. *Nucleic Acids Res*., **31**, 23–27.
Fischer,M. and Pleiss,J. (2003) The Lipase Engineering Database: a navigation and analysis tool for protein families. *Nucleic Acids Res*., **31**, 319–321.
Fischer,M. *et al*. (2006) DWARF – a data warehouse system for analyzing protein families. *BMC Bioinformatics*, **7**, 495.
Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.
Kemper,B. (2004) Structural basis for the role in protein folding of conserved proline-rich regions in cytochromes P450. *Toxicol Appl. Pharmacol*., **199**, 305–315.
Lisitsa,A.V. *et al*. (2001) Cytochrome P450 database. *SAR QSAR Environ Res*., **12**, 359–366.
Mestres,J. (2005) Structure conservation in cytochromes P450. *Proteins*, **58**, 596–609.
Montellano,O.d. (1995) Cytochrome P450: structure, mechanism and biochemistry. New York, Plenum Press.
Nelson,D.R. (2006) Cytochrome P450 nomenclature, 2004. *Methods Mol. Biol*., **320**, 1–10.
Nelson,D.R. *et al*. (2002) Mining databases for cytochrome P450 genes. *Methods Enzymol*., **357**, 3–15.
Raucy,J.L. and Allen,S.W. (2001) Recent advances in P450 research. *Pharmacogenomics J*, **1**, 178–186.
Rice,P. *et al*. (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet*., **16**, 133–154.
Schwede,T. *et al*. (2000) Protein structure computing in the genomic era. *Res Microbiol*., **151**, 107–112.
Thompson,J.D. *et al*. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*., **22**, 4673–4680.