Uncovering of Cytochrome P450 Anatomy by SecStrAnnotator

Adam Midlik[1,2], Veronika Navrátilová[3,†], Taraka Ramji Moturu[1,2], Jaroslav Koča[1,2], Radka Svobodová[1,2,*] and Karel Berka[3,*]

[1]CEITEC - Central European Institute of Technology, Masaryk University, Kamenice 5, 625 00 Brno, Czech Republic
[2]National Centre for Biomolecular Research, Faculty of Science, Masaryk University, Kamenice 5, 625 00 Brno, Czech Republic
[3]Department of Physical Chemistry, Regional Centre of Advanced Technologies and Materials, Faculty of Science, Palacký University, 17. listopadu 1192/12, 771 46 Olomouc, Czech Republic
*To whom correspondence should be addressed.
†Current address: Pharmacovigilance department, State Institute for Drug Control, Šrobárova 48, 100 41 Praha 10, Czech Republic

Abstract

Motivation: Protein structural families are groups of homologous proteins defined by the organization of secondary structure elements (SSEs). Nowadays, many families contain vast numbers of homologous structures and the SSEs can help to orient within them. Communities around specific protein families have even developed specialized SSE annotations, assigning always the same name to the equivalent SSEs in homologous proteins. A detailed analysis of the groups of equivalent SSEs and their variability provides an overview of the studied protein family and can be used to enrich the analysis of a particular protein at hand.
Results: We developed a workflow for analysis of the secondary structure anatomy of a protein family, based on SSE annotation tool SecStrAnnotator. We applied this analysis to the model family of cytochromes P450 (CYPs) – a family of important biotransformation enzymes with a community-wide used SSE annotation. We report the occurrence, typical length and amino acid sequence for the equivalent SSE groups, as well as the conservation/variability of these properties. We also suggest a generic residue numbering scheme for the CYP family. The comparison between the bacterial and eukaryotic part of the family highlights the major differences and reveals an anomalous group of bacterial CYPs with some typically eukaryotic features. This hints at possible evolutionary and functional relationships.
Availability: The software and data are available at https://webchem.ncbr.muni.cz/Wiki/SecStrAnnotator
Contact: radka.svobodova@ceitec.muni.cz, karel.berka@upol.cz

1 Introduction

Secondary structure elements (SSEs) are defined by the repetitive pattern of hydrogen bonds and geometric arrangement. The most well-known SSE types are the α-helix and the β-strand. SSEs have been used to analyze protein structures since their first observation by Linus Pauling (Pauling *et al.*, 1951; Pauling and Corey, 1951). They define the structural folds of individual protein structural families as classified by *CATH* (Dawson *et al.*, 2017) or *SCOPe* (Fox *et al.*, 2014) databases. SSEs play an important role in the determination of the function of those protein families (Thornton *et al.*, 2000). SSEs can also serve as guides for orientation in the protein structures within scientific communities.

In order to compare similar structures, the communities around several proteins families have developed specialized nomenclatures for annotation (labeling) of the SSEs in the members of the family. Such nomenclatures assign the same label to the equivalent SSEs from different proteins in the family. We will refer to a group of equivalent SSEs as an *SSE class*.

Adoption of such SSE nomenclatures is typical for well-studied protein families with a large number of available structures with a common fold but large sequence variations, such as esterases (Krejci *et al.*, 1991; Lenfant *et al.*, 2013), G-protein coupled receptors (GPCRs) (Isberg *et al.*, 2015), immunoglobulins (Ehrenmann *et al.*, 2010), cytochromes P450 (CYPs) (Rowland *et al.*, 2006) and others. These traditional nomenclatures prove to be particularly useful when comparing existing structures, describing new ones or generalizing observations over the whole family.

Annotated SSEs can be used as reference points to describe the position of key regions, such as catalytic sites, selectivity filters, channels or protein-protein interfaces. A nice illustrative example is again the CYP family (Fig. 1), with a well-established classification of multiple different channels based on their position relative to the annotated SSEs (Cojocaru *et al.*, 2007; Hendrychova *et al.*, 2012; Yu *et al.*, 2013) and with selectivity defining residues on several SSEs (Otyepka *et al.*, 2007; Otyepka *et al.*, 2012; Urban *et al.*, 2018).

In some protein families, the communities have extended the annotation down to the residue level and introduced generic numbering schemes. Such schemes assign the same generic number to the equivalent residue positions in homologous proteins, which facilitates comparisons of mutation effects, ligand interactions, structural motifs etc. The generic numberings may be based on the sequence information (e.g. immunoglobulins (Dunbar and Deane, 2016)) or may combine the SSE annotations with the sequence information (e.g. GPCRs (Isberg *et al.*, 2015)).

Annotation of SSEs can be valuable even in protein families without defined traditional nomenclature since it provides the equivalence between the SSEs from the individual members of the family (the SSEs from the same SSE class are annotated by the same label, even if the labels are arbitrarily created).

Previously, we presented methods for automated annotation of SSEs in protein families, implemented in tool SecStrAnnotator (Midlik *et al.*, 2019).

Automated annotation opens the possibility to focus on any protein family and describe its general SSE anatomy, which can bring a valuable insight to the understanding of its function – SSEs typically present, their occurrence, typical length, position and amino acid composition and variation for individual SSE classes.

In this paper, we propose a procedure for analysis of the general SSE anatomy of a protein structural family, based on SecStrAnnotator. We demonstrate this type of analysis on the cytochromes P450, a biologically important family with a long tradition of SSE annotation and well-established SSE nomenclature. We also suggest a generic residue numbering scheme for the CYP family. The SSE annotations for the CYP family are accessible online through SecStrAPI and can be easily visualized via a dedicated PyMOL plugin – all freely available at https://webchem.ncbr.muni.cz/Wiki/SecStrAnnotator.

1.1 Traditional SSE nomenclature in the CYP family

The common fold of the CYP family has a triangular prism shape and consists mostly of α-helices combined with several β-sheets and a heme cofactor, which forms the catalytic center of the enzyme (Peterson and Graham, 1998; Otyepka *et al.*, 2007; Johnson and Stout, 2013). The traditional SSE nomenclature in the CYP family is based on the labels used by Poulos (Poulos *et al.*, 1985) for the first experimentally determined CYP structure (P450cam) with 12 helices, labeled A–L, and 5 sheets, β1–β5. The publication of the refined structure (Poulos *et al.*, 1987) mentioned a new helix B′ and also several shorter helices and strands without labels.

Ravichandran *et al.* (1993) on P450 BM3 changed the labeling scheme for β-sheets to the form which later became widely used in the community: sheets β1 (5 strands, previously labeled β1+β3), β2 (2 strands, previously β4), β3 (3 strands, previously β5) and a new sheet β4 (2 strands). The strands within each sheet can be referred to individually, using a hyphen (e.g. β1-1, β1-2). They also added annotation of two new helices – J′ (between J and K) and K′ (after strand β1-3).

As more and more structures emerged in the following years, new labels were needed for the newly observed SSE classes: helices A′ (Hasemann *et al.*, 1994), L′ (Park *et al.*, 1997; Scott *et al.*, 2003), F′, G′ (Wester *et al.*, 2003), K″ (Scott *et al.*, 2003) and B″ (Ouellet *et al.*, 2008); sheets β5 (Hasemann *et al.*, 1994) (corresponding to β2 in (Poulos *et al.*, 1987)) and β6 (Hasemann *et al.*, 1995); and strand β1-0 (Polypenko *et al.*, 2003). Many other SSEs have been mentioned and labeled in literature but these are either very rare or can be treated as a part of a longer SSE (e.g. helix D′ in (Park *et al.*, 1997) can be understood as an N-terminal part of helix D).

Unfortunately, the labeling is not always consistent, sometimes even in the papers by the same author. The same SSE class can be assigned different labels (e.g. helix L′ in (Scott *et al.*, 2003) is helix M in (Polypenko *et al.*, 2003)) or one label can be assigned to different SSE classes (e.g. helix A″ is located between β1-1 and β1-2 in (Polypenko *et al.*, 2003) but before A′ in (Williams *et al.*, 2004)). Throughout this paper we will use the nomenclature as shown in Fig. 1 and specified in Supplementary Table S1.

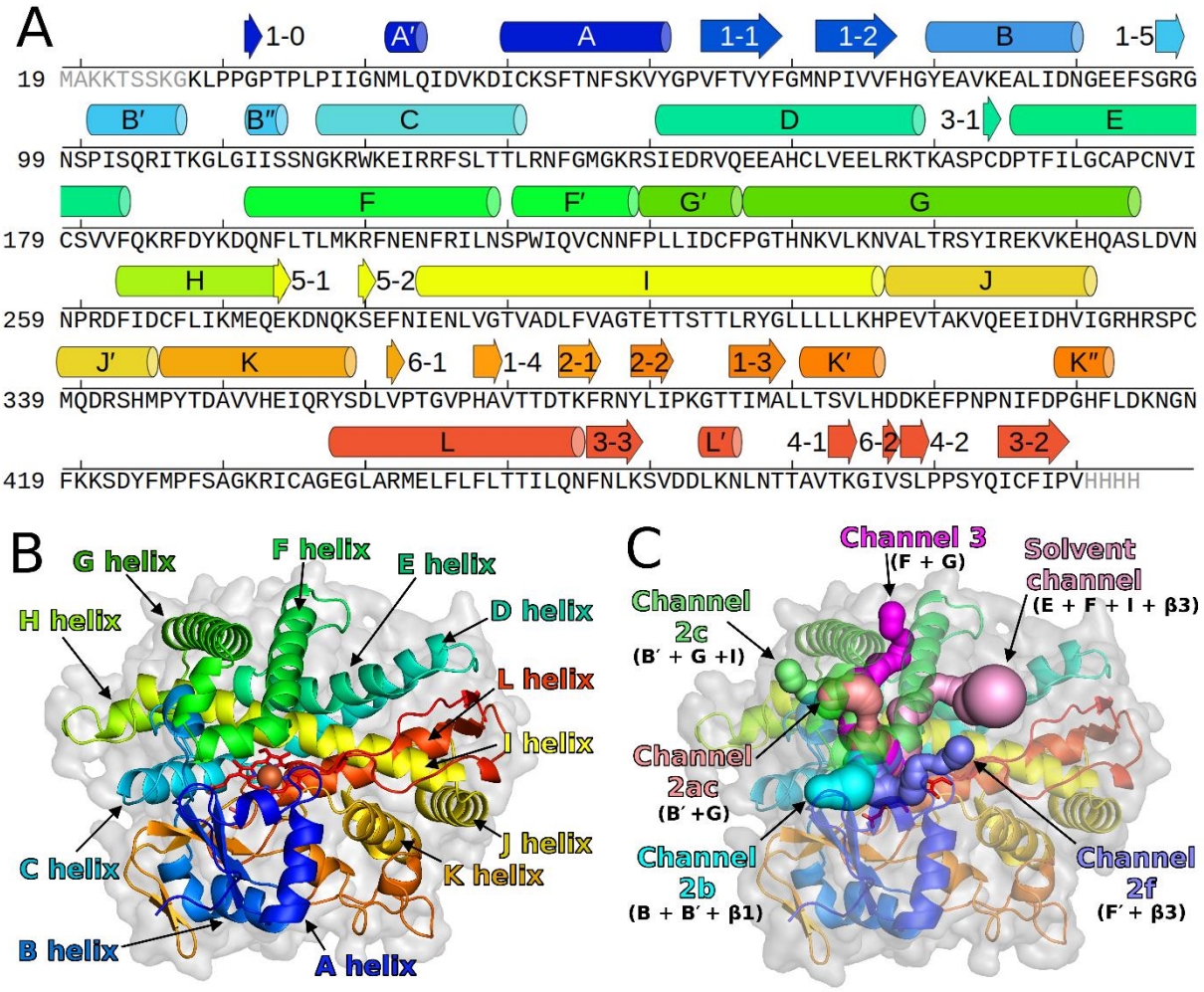

Fig. 1. SSE annotation for the CYP family. A) annotation shown on the sequence (PDB ID 2nnj), B) annotation of the major helices shown on the structure, C) annotation of protein channels defined with respect to the annotated SSEs. Some SSEs (B″, β1-0, β5-1, β5-2, β6-1, β6-2) are not really present in 2nnj but they are shown to illustrate their location in other CYPs. The table of residue ranges of the SSEs is included in Supplementary Table S1.

2 Methods

2.1 Datasets

Set-NR: A list of protein domains annotated as Cytochrome P450 was obtained from SIFTS resource (Dana *et al.*, 2019) via PDBe REST API (available at http://www.ebi.ac.uk/pdbe/api/mappings/:accession) on 23 March 2020. More specifically, annotations originating from databases CATH and Pfam (accessions 1.10.630.10 and PF00067) were merged to obtain 1807 protein domains located in 973 PDB entries. The information about residue ranges was

discarded and whole chains were taken instead (this was necessary because Pfam often wrongly annotates only a small portion of the chain). The domains were mapped to UniProt IDs and the best-quality domain was selected for each UniProt ID. The quality was measured by "overall_quality" obtained from PDBe REST API (http://www.ebi.ac.uk/pdbe/api/validation/summary_quality_scores/entry/:pdbid). The domains which map to no UniProt ID were excluded. The resulting Set-NR (non-redundant) contains 175 protein domains.

Set-NR-Bact and Set-NR-Euka: Domains from Set-NR were mapped to their source organism (http://www.ebi.ac.uk/pdbe/api/pdb/entry/molecules/:pdbid) and divided into four subsets based on their superkingdom using NCBI Taxonomy (Sayers *et al.*, 2009): Set-NR-Bact (Bacteria, 120 structures), Set-NR-Euka (Eukaryota, 51 structures), Set-NR-Arch (Archaea, 3 structures) and Set-NR-Viru (Viruses, 1 structure). However, Set-NR-Arch and Set-NR-Viru were not analyzed separately because of their small size.

Lists of PDB IDs and UniProt IDs for each dataset are provided as a TSV file in the Supplementary data.

2.2 Template annotation

Since SecStrAnnotator requires an annotated template structure, we have chosen a template domain based on multiple selection criteria.

First, the template should contain all SSE classes. Thus, we considered only the eukaryotic structures (Set-NR-Euka).

Second, the template structure should be an "average" structure which is as similar to all the others as possible. Therefore, we compared each pair of structures in Set-NR-Euka by *cealign* command in PyMOL and calculated pairwise Q-scores (Krissinel and Henrick, 2004). Then for each structure we calculated the average Q-score against all the other structures Q_{avg}. We selected the structure with the highest Q_{avg} as the template domain, which was 2nnjA (human CYP 2C8).

Third, template structure should have sufficient resolution, quality and should not contain unmodeled loops etc. The selected domain 2nnjA meets these criteria (resolution 2.28 Å, overall quality 42.46, observed residue range 10–472 covers the whole region of interest).

Secondary structure annotation was mapped from the annotation of CYP 2C9 by Rowland *et al.* (2006) with several added SSE classes, as described in Section 1.1 and shown in Fig. 1A.

2.3 Annotation procedure

The annotation was performed using our software SecStrAnnotator (version 2.1). The algorithm consists of three main steps: secondary structure assignment, structural alignment and SSE matching. Detailed description is provided in (Midlik *et al.*, 2019). SecStrAnnotator was run with these settings: --ssa geom-hbond2 --align cealign --matching mom --soft --maxmetric 25,0.5,0.5 --label2auth --verbose.

2.4 Statistical evaluation

All statistical tests were performed in R (version 3.4.4-1ubuntu1) using the *stats* library. The plots were generated in R using the *ggplot2* library.

Comparison of bacterial and eukaryotic dataset: The occurrence of each SSE class in Set-NR-Bact and Set-NR-Euka was modeled as a binomial distribution, and the two datasets were compared by the test of equal proportions (*prop.test*) with $\alpha = 0.05$.

The distribution of length (number of residues) of each SSE class was compared between Set-NR-Bact and Set-NR-Euka. Where the medians of the eukaryotic and bacterial distribution were not equal, the two-sample Kolmogorov-Smirnov test (*ks.test*) with $\alpha = 0.05$ was used to decide if the difference between the distributions is significant. Non-existing SSEs were not included in the length distributions.

2.5 Multiple sequence alignment

The amino acid sequences of the individual SSE classes were extracted from Set-NR and aligned using an in-house algorithm NoGapAligner which allows gaps only at the beginning and at the end but not within a sequence (this is necessary in order to establish generic residue numbering). Substitution matrix was BLOSUM62 and the gap penalty was set to 10. Sequence logos were produced using *logomaker* module for Python (Tareen and Kinney, 2019).

For every position i in each multiple sequence alignment, the conservation measure was calculated as $c_i = p_i \times I_i$, where p_i is the fraction of sequences not having a gap on the position i and I_i is the information content of the position i (Schneider and Stephens, 1990). (In the visual form of a logo, c_i corresponds to the area of the i-th column.) The position with the greatest c_i within the alignment was selected as the reference residue of the SSE class.

3 Results

Structures of proteins from the CYP family typically contain at least 14 helices – A, B, B′, C, D, E, F, G, H, I, J, K, K′, L – and 4 sheets – β1 (5 strands), β2 (2 strands), β3 (3 strands), β4 (2 strands). Additional helices A′, B″, F′, G′, J′, K″ and L′ are often present, as well as two sheets β5 and β6 (with two strands in each). Sheet β1 often contains an extra strand β1-0 (bonded to β1-1). The sequential order and 3D position of these SSE classes is shown in Fig. 1. Other SSEs may appear in the structures, but they are not characteristic to the family.

For convenience, we divide all annotated SSE classes into three groups throughout the paper: *major helices* (helices A–L, typically longer than 8 residues), *minor helices* (all the remaining helices, typically shorter than 8 residues) and *strands*.

In the following text, we analyze these SSE classes in terms of frequency of occurrence and amino acid composition. We also discuss differences between eukaryotic and bacterial structures. In the Supplementary data we provide a dedicated analysis of the lengths and structural irregularities.

3.1 Frequency of occurrence of the SSE classes

This section describes the frequency of occurrence of each SSE class, i.e. in what fraction of the structures the particular SSE is present. The results are shown in Fig. 2.

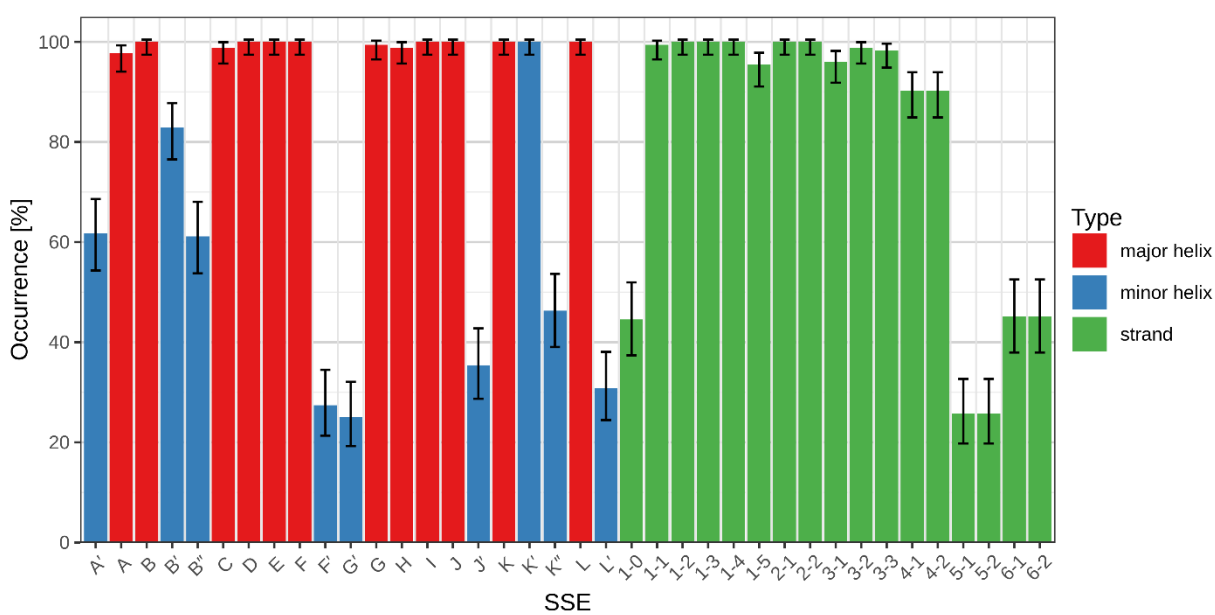

Fig. 2. Frequency of occurrence of individual SSE classes. Error bars show confidence intervals calculated by the Agresti-Coull method for α = 0.05 (Agresti and Coull, 1998).

Major helices

All major helices occur in more than 97% of the structures. In cases where they are missing, it can be attributed to the experiment (bad quality, residue coverage or resolution of the structure), rather than not being formed in the structure.

Minor helices

The most frequent minor helices are K′ (100%) and B′ (83%), followed by A′ (62%), B″ (61%), K″ (46%), J′ (35%), L′ (31%), F′ (27%) and G′ (25%). Helices A′, B″, K″, L′ are very short, so it can often happen that they are not formed at all. On the other hand, the low occurrence of helices F′, G′, J′ can be explained by their absence in bacterial CYPs (roughly 2/3 of all structures) – see Section 3.3 for more details. Furthermore, the flexible F′G′-loop is often not modeled in the experimental structures.

Strands

In sheet β1, strands β1-1, β1-2, β1-3, β1-4 are always present, β1-5 is found in 95% of the structures, β1-0 in 45%. Strand β1-0 is less frequent because it is very short, and it is in the region of variable secondary structure before helix A. Sheet β2 is always present. In sheet β3, strands β3-2, β3-3 are present in more than 98% of structures (if they are missing, it is because the corresponding residues are not modeled in the experimental structure). β3-1 is sometimes not formed (present in 96%). Sheet β4 is found in 90% of structures – sometimes it is not formed. The remaining two sheets are much less frequent – β5 (26%) and β6 (45%).

3.2 Multiple sequence alignment and generic residue numbering

For SSE classes with sufficient sequence conservation we created sequence logos (see Fig. 3) and selected the most conserved residue as the reference residue. Based on the reference residue, we established a generic residue numbering scheme similar to the schemes used for GPCRs, described by Isberg *et al.* (2015) and used throughout GPCRdb (Pándy-Szerekes *et al.*, 2017), or immunoglobulins (Dunbar and Deane, 2016).

The reference residue is always numbered as @X.50, where X is the SSE label (character @ is added to avoid confusion in line notation). The remaining residues are then numbered correspondingly. An example of such residue identification is W120$^{@C.46}$ in structure 2nnj (or generically @C.46), denoting the tryptophan residue 120 in helix C four positions before the reference residue, which is R124$^{@C.50}$. Residue mutations can be also specified, e.g. W120A$^{@C.46}$ (Fig. 3).

We established the generic numbering for those SSEs, that contain at least one column in their logo with area (c_i) greater than 2 bits: major helices B, C, E, H, I, J, K, L, minor helices J′, K′, K″, and strands β1-1, β1-2, β1-3, β1-4, β1-5, β2-2, β3-1, β3-2.

Other SSE classes have insufficient sequence conservation and/or sequence alignment does not correspond to structure alignment (i.e. reference residues do not align in 3D). Therefore, it is impossible to establish a meaningful generic numbering for these SSE classes. All sequence logos are available in the Supplementary data.

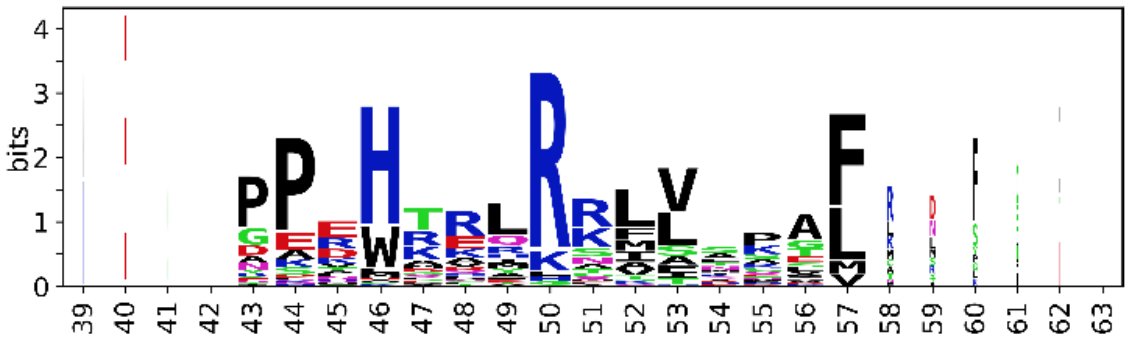

Fig. 3. Sequence logo of helix C with generic numbering of residues. The remaining sequence logos are available in the Supplementary data.

3.3 Comparison of bacterial and eukaryotic CYPs

We compared the frequencies of occurrence of individual SSE classes between bacterial and eukaryotic structures (Fig. 4) as well as their distributions of length (Fig. 5).

Helices A′, B′, F′, G′, J′ and K″ are significantly more frequent in eukaryotes. This is in agreement with literature (Otyepka *et al.*, 2007; Johnson and Stout, 2013) reporting that F′ and G′ are typically not present in bacteria. However, sometimes a short helix can be formed between F and G in bacterial structures, which is then automatically annotated as F′ or G′, so the observed occurrence is not equal to zero even in bacteria.

Conversely, helix B″, sheet β6 and strand β1-5 occur more frequently in bacteria.

Helices A, A′, C, F, G, H, J, J′ and strands β1-1, β1-2, β3-2, β3-3, β6-2 are longer in eukaryotes, while strand β1-4 tends to be longer in bacteria. The case of β6-2 can be explained by the merging of β6-2 and β4-2 in some eukaryotic structures. The Kolmogorov-Smirnov test reports significant length difference for a few more SSE classes (B, K″ towards eukaryotes; D, K, L, β1-0, β1-5 towards bacteria), however in these cases the mean length differs by less than one residue.

A notable difference is visible in the region of J and J′ helices. Their typical length in eukaryotes is 15 and 7 residues, respectively, while in bacteria helix J has typically 10 residues and J′ is not present at all. However, there is an anomalous group of 5 bacterial CYPs whose helices J and J′ have lengths 15 and 7 residues, exactly as observed in eukaryotic CYPs. This group includes:

- the heme domain of the flavocytochrome P450 BM3 (*Bacillus megaterium*, PDB IDs 3kx3, 6h1l and 6h1t (these map to different UniProt IDs but have the same sequence of the heme domain)), where helix J' plays a role in the interaction with redox flavodomain,
- CYP 51 (*Mycobacterium tuberculosis*, PDB ID 2ci0),
- CYP 51 (*Methylococcus capsulatus*, PDB ID 6mcw),
- CYP 120A1 (*Synechocystis sp.*, PDB ID 3ve3),
- CYP 170A1 (*Streptomyces coelicolor*, PDB ID 3dbg).

We further investigated this anomalous bacterial group and discovered more similarities to the eukaryotic structures:

- The sequence of their helix J resembles the sequence of eukaryotic helix J – most notably the glutamic acid E@J.59 (see Section 3.2) is highly conserved in both the eukaryotic CYPs (98%) and the anomalous group (100%) due to interactions with backbone amide groups of residues @K.40 and @K.41, while being much less conserved in the rest of the bacterial CYPs (7%), where this interaction is not observed.
- The region between helix K′ and the heme binding site is approximately 9 residues longer in the eukaryotic and anomalous CYPs compared to the regular bacterial CYPs (roughly 39 residues in the eukaryotic and anomalous CYPs, 30 residues in the regular bacterial CYPs). Furthermore, in eukaryotic and anomalous CYPs, this region typically contains helix K″ (in 84% of the eukaryotic and in all anomalous CYPs), which is usually not present in the regular bacterial CYPs (only in 27%).

These deviations are concentrated in the region which has been reported as the interface for binding the redox partner of the CYP (for P450 BM3 (Sevrioukova *et al.*, 1999), human mitochondrial CYP11A1 (Strushkevich *et al.*, 2011)).

In some other aspects the anomalous CYPs behave as typical bacterial CYPs – there is no F′ and G′ helix in the FG-loop; the A-propionate side chain of the heme is oriented to the distal side (towards the substrate binding pocket). We can therefore hypothesize that this group represents evolutionary transition towards eukaryotic CYPs – this is also supported by the fact that the anomalous bacterial CYPs group with the eukaryotic sequences in the phylogenetic tree from Set-NR (Supplementary Fig. S6).

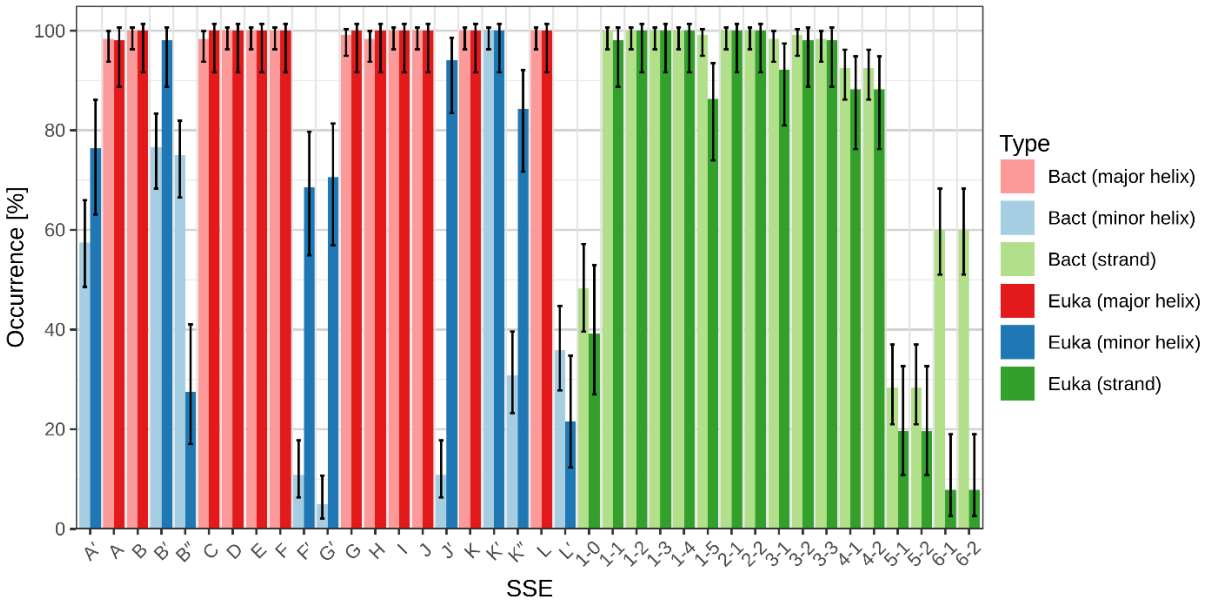

Fig. 4. Comparison of SSE class occurrence in bacterial (Bact) and eukaryotic (Euka) CYP structures.

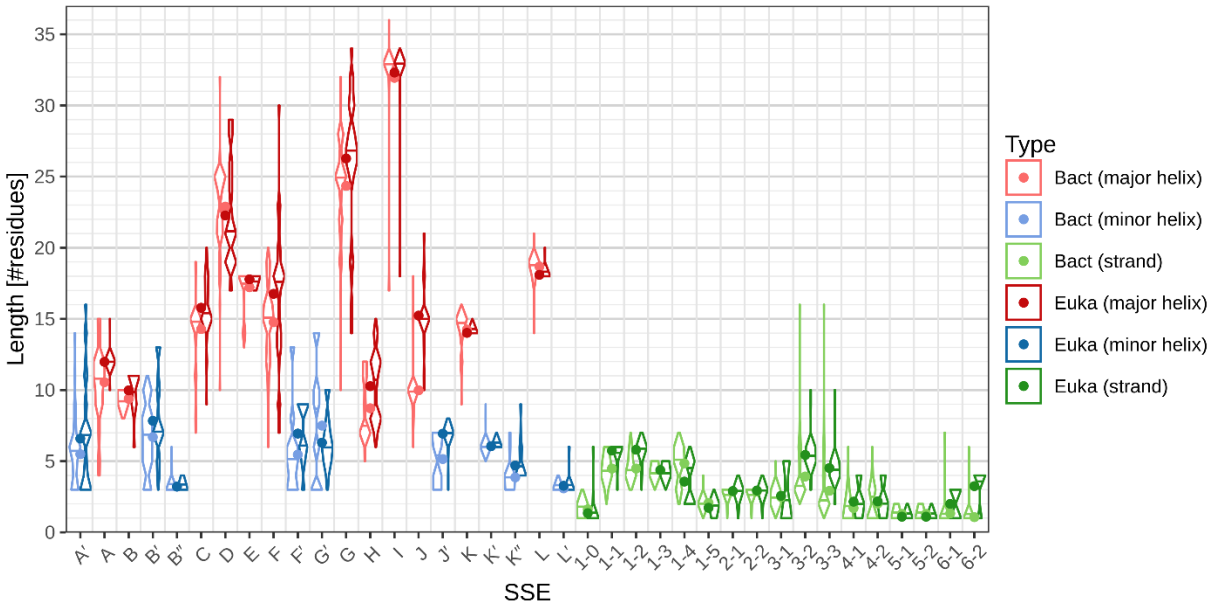

Fig. 5. Comparison of SSE class length distribution in bacterial (Bact) and eukaryotic (Euka) CYP structures. The absent SSEs are not included in the distributions.

3.4 SecStrAPI: how to get to our annotations

All annotations which are mentioned in this paper are publicly available through SecStrAPI (https://webchem.ncbr.muni.cz/API/SecStr).

The annotations can be downloaded directly (in JSON format, described in detail on the website) or can be accessed through PyMOL plugin *secstrapi_plugin.py*, which is available on the website and serves for simple and quick visualization of the SSE annotations.

4 Conclusion

We presented a workflow for description of the secondary structure anatomy of a protein structural family – automatic annotation of secondary structure elements, analysis of their frequency of occurrence, typical length, position, amino acid composition and the variability of these properties.

These methods were demonstrated in the case study of the Cytochromes P450 (CYP) family. The characteristic SSEs of the family are 14 helices A, B, B′, C, D, E, F, G, H, I, J, K, K′, L and 4 sheets β1, β2, β3, β4, which occur nearly in all structures. Optional SSEs include helices A′, B″, F′, G′, J′, K″, L′, sheets β5, β6, and strand β1-0. Some of these SSE classes are very uniform in length (the core helices C, E, I, K, L, but also J′, K′), while some show extensive length variation (A′, A, B′, D, F, F′, G′, G, H, J, K″, β3-2, β3-3). The shortest helices B″, L′ and sheets β5, β6 are on the edge of detection.

For the SSE classes with sufficient sequence conservation we have established a generic residue numbering scheme (e.g. W120$^{@C.46}$) similar to that used for the GPCR family. However, this was not possible in the variable regions (A, B′, D, F, G), which are responsible for substrate uptake and recognition.

We also compared the eukaryotic and bacterial members of the CYP family. The most significant difference is the absence of helices F′, G′, J′ in bacteria. Helices A′, B′, K″ are also rarer in bacteria, while B″, β6 and β1-5 are more common in bacteria than in eukaryotes. Many SSEs tend to be longer in either eukaryotes (A, A′, F, G, H, J, J′, β1-1, β1-2, β3-2, β3-3, β6-2) or bacteria (β1-4).

Strikingly, we also identified a small group of 5 bacterial CYPs with typical eukaryotic features in the region of helices J–L, which can be explained by the interaction with the redox partner.

All the utilized software tools and the obtained data, including secondary structure annotations and generic residue numbers, are available at https://webchem.ncbr.muni.cz/Wiki/SecStrAnnotator. The annotations can be easily visualized with a PyMOL plugin. While the annotations are now only specific for the CYP family, the software is applicable in principle to all protein families with defined annotations. We are currently working on the generalization of the annotation pipeline over all CATH protein families.

Acknowledgements

We would like to thank Veronika Bendová for her valuable advice on the statistical procedures.

Funding

This work was supported by the Ministry of Education, Youth and Sports of the Czech Republic under the project CEITEC 2020 [LQ1601]; ELIXIR-CZ research infrastructure project including access to computing and storage facilities [LM2018131]; European Regional Development Fund – projects ELIXIR-CZ [CZ.02.1.01/0.0/0.0/16_013/0001777]; Nanotechnologies for Future [CZ.02.1.01/0.0/0.0/16_019/0000754]; and Brno City Municipality - Brno Ph.D. Talent Scholarship to A.M..

Conflict of interest

None declared.

References

Agresti,A. and Coull,B.A. (1998) Approximate is better than 'exact' for interval estimation of binomial proportions. *Am. Stat.*, **52**, 119–126.

Cojocaru,V. *et al*. (2007) The ins and outs of cytochrome P450s. *Biochim. Biophys. Acta-Gen. Subj.*, **1770**, 390–401.

Dana,J.M. *et al*. (2019) SIFTS: updated Structure Integration with Function, Taxonomy and Sequences resource allows 40-fold increase in coverage of structure-based annotations for proteins. *Nucleic Acids Res.*, **47**, D482–D489.

Dawson,N.L. *et al.* (2017) CATH: an expanded resource to predict protein function through structure and sequence. *Nucleic Acids Res.*, **45**, D289–D295.

Dunbar,J. and Deane,C.M. (2016) ANARCI: antigen receptor numbering and receptor classification. *Bioinformatics*, **32**, 298–300.

Ehrenmann,F. *et al.* (2010) IMGT/3Dstructure-DB and IMGT/DomainGapAlign: a database and a tool for immunoglobulins or antibodies, T cell receptors, MHC, IgSF and MhcSF. *Nucleic Acids Res.*, **38**, D301–7.

Fox,N.K. *et al.* (2014) SCOPe: Structural Classification of Proteins—extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Res.*, **42**, D304–309.

Hasemann,C.A. *et al.* (1994) Crystal structure and refinement of cytochrome P450terp at 2.3 A resolution. *J. Mol. Biol.*, **236**, 1169–85.

Hasemann,C.A. *et al.* (1995) Structure and function of cytochromes P450: a comparative analysis of three crystal structures. *Structure*, **3**, 41–62.

Hendrychova,T. *et al.* (2012) Dynamics and hydration of the active sites of mammalian cytochromes P450 probed by molecular dynamics simulations. *Curr. Drug Metab.*, **13**, 177–89.

Hutchinson,E.G. and Thornton,J.M. (1990) HERA - A program to draw schematic diagrams of protein secondary structures. *Proteins*, **8**, 203–212.

Hutchinson,E.G. and Thornton,J.M. (1996) PROMOTIF - A program to identify and analyze structural motifs in proteins. *Protein Sci.*, **5**, 212–220.

Isberg,V. *et al.* (2015) Generic GPCR residue numbers - aligning topology maps while minding the gaps. *Trends Pharmacol. Sci.*, **36**, 22–31.

Johnson,E.F. and Stout,C.D. (2013) Structural diversity of eukaryotic membrane cytochrome P450s. *J. Biol. Chem.*, **288**, 17082–90.

Katoh,K. and Standley,D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.*, **30**, 772–80.

Krejci,E. *et al.* (1991) Cholinesterase-like domains in enzymes and structural proteins: functional and evolutionary relationships and identification of a catalytically essential aspartic acid. *Proc. Natl. Acad. Sci. U. S. A.*, **88**, 6647–51.

Krissinel,E. and Henrick,K. (2004) Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr. Sect. D-Struct. Biol.*, **60**, 2256–68.

Lenfant,N. *et al.* (2013) ESTHER, the database of the α/β-hydrolase fold superfamily of proteins: tools to explore diversity of functions. *Nucleic Acids Res.*, **41**, D423–9.

Midlik,A. *et al.* (2019) Automated Family-Wide Annotation of Secondary Structure Elements. In Kister,A.E. (ed.), *Protein Supersecondary Structures. Methods in Molecular Biology*, vol 1958. Humana Press, New York, pp. 47–71.

Otyepka,M. *et al.* (2007) What common structural features and variations of mammalian P450s are known to date? *Biochim. Biophys. Acta-Gen. Subj.*, **1770**, 376–89.

Otyepka,M. *et al.* (2012) Is there a relationship between the substrate preferences and structural flexibility of cytochromes P450? *Curr. Drug Metab.*, **13**, 130–42.

Ouellet,H. *et al.* (2008) Mycobacterium tuberculosis CYP130: crystal structure, biophysical characterization, and interactions with antifungal azole drugs. *J. Biol. Chem.*, **283**, 5069–80.

Pándy-Szekeres,G. *et al.* (2018) GPCRdb in 2018: adding GPCR structure models and ligands. *Nucleic Acids Res.*, **46**, D440–D446.

Park,S.Y. *et al.* (1997) Crystal structure of nitric oxide reductase from denitrifying fungus Fusarium oxysporum. *Nat. Struct. Biol.*, **4**, 827–32.

Pauling,L. *et al.* (1951) The Structure of Proteins - Two Hydrogen-Bonded Helical Configurations of the Polypeptide Chain. *Proc. Natl. Acad. Sci. U. S. A.*, **37**, 205–211.

Pauling,L. and Corey,R.B. (1951) Configurations of Polypeptide Chains With Favored Orientations Around Single Bonds: Two New Pleated Sheets. *Proc. Natl. Acad. Sci. U. S. A.*, **37**, 729–40.

Peterson,J.A. and Graham,S.E. (1998) A close family resemblance: the importance of structure in understanding cytochromes P450. *Structure*, **6**, 1079–85.

Pylypenko,O. *et al.* (2003) Crystal structure of OxyC, a cytochrome P450 implicated in an oxidative C-C coupling reaction during vancomycin biosynthesis. *J. Biol. Chem.*, **278**, 46727–33.

Poulos,T.L. *et al.* (1985) The 2.6-A Crystal Structure of Pseudomonas putida Cytochrome P-450. *J. Biol. Chem.*, **260**, 16122–16130.

Poulos,T.L. *et al.* (1987) High-resolution crystal structure of cytochrome P450cam. *J. Mol. Biol.*, **195**, 687–700.

Ravichandran,K.G. et al. (1993) Crystal structure of hemoprotein domain of P450BM-3, a prototype for microsomal P450's. *Science*, **261**, 731-6.

Rowland,P. *et al.* (2006) Crystal structure of human cytochrome P450 2D6. *J. Biol. Chem.*, **281**, 7614–7622.

Sayers,E.W. *et al.* (2009) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **37**, D5–15.

Schneider,T.D. and Stephens,R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, **18**, 6097–6100.

Scott,E.E. *et al.* (2003) An open conformation of mammalian cytochrome P450 2B4 at 1.6-A resolution. *Proc. Natl. Acad. Sci. U. S. A.*, **100**, 13196–201.

Sevrioukova,I.F. *et al.* (1999) Structure of a cytochrome P450-redox partner electron-transfer complex. *Proc. Natl. Acad. Sci. U. S. A.*, **96**, 1863–8.

Strushkevich,N. *et al.* (2011) Structural basis for pregnenolone biosynthesis by the mitochondrial monooxygenase system. *Proc. Natl. Acad. Sci. U. S. A.*, **108**, 10139–43.

Tareen,A. and Kinney,J.B. (2019) Logomaker: beautiful sequence logos in Python. *Bioinformatics*, btz921.

Thornton,J.M. *et al.* (2000) From structure to function: approaches and limitations. *Nat. Struct. Biol.*, **7**, 991–4.

Urban,P. *et al.* (2018) Ligand Access Channels in Cytochrome P450 Enzymes: A Review. *Int. J. Mol. Sci*., **19**, 1617.

Wester,M.R. *et al.* (2003) Structure of a substrate complex of mammalian cytochrome P450 2C5 at 2.3 A resolution: evidence for multiple substrate binding modes. *Biochemistry*, **42**, 6370–9.

Williams,P.A. *et al.* (2004) Crystal structures of human cytochrome P450 3A4 bound to metyrapone and progesterone. *Science*, **305**, 683–6.

Yu,X. *et al.* (2013) Conformational diversity and ligand tunnels of mammalian cytochrome P450s. *Biotechnol. Appl. Biochem.*, **60**, 134–45.