Chapter 2

Effective Techniques for Protein Structure Mining

Stefan J. Suhrer, Markus Gruber, Markus Wiederstein, and Manfred J. Sippl

Abstract

Retrieval and characterization of protein structure relationships are instrumental in a wide range of tasks in structural biology. The classification of protein structures (COPS) is a web service that provides efficient access to structure and sequence similarities for all currently available protein structures. Here, we focus on the application of COPS to the problem of template selection in homology modeling.

Key words: Protein structure space, Protein structure comparison, Template selection, Structure alignment, Structure similarity search, Classification, Homology modeling, Ligand binding

1. Introduction

The repository of known protein structures contains a wealth of information about the relationships between protein sequences and protein structures. Many useful tools and databases have been developed to extract knowledge from this repository, but the appropriate organization of protein structure data remains a challenge.

The classification of protein structures (COPS) (1–3) provides access to the overwhelming number of structure and sequence relationships (4, 5) between all experimentally determined protein structures deposited in the Protein Data Bank (PDB) (6). COPS features a quantitative organization of protein structures according to a set of metric properties and principles. It includes methods for the automated decomposition of proteins into structural domains, pairwise structure comparison, and the instant visualization of structure similarities. Since COPS is updated weekly with every PDB release, it covers the complete set of publicly available protein structures.

Andrew J.W. Orry and Ruben Abagyan (eds.), *Homology Modeling: Methods and Protocols*, Methods in Molecular Biology, vol. 857, DOI 10.1007/978-1-61779-588-6_2,

In this chapter, we present and illustrate the usage of COPS with an emphasis on its use in homology modeling. Homology modeling builds on the observation that proteins of similar sequence frequently adopt similar structures (7). Proteins of unknown structure are modeled using the structures of other proteins as templates, given their sequences share significant similarity. In this procedure, the steps of template selection, template comparison, and evaluation for their use in model building are significantly affected by the way protein structure data is organized and accessible. Moreover, it is important to keep pace with the rapid growth of PDB which implies an ever increasing pool of template candidates. We discuss the key components of COPS and apply them to the step of template characterization in homology modeling.

2. Structure Mining with COPS

The COPS classification process includes the weekly download of structures from PDB, their decomposition into domains with TopDomain, the calculation of structural similarities with TopMatch (8), and the update of the COPS hierarchy with respect to the found similarities. The domains are organized in a tree similar to a file browser, where the domains correspond to tree nodes and pairwise structural similarities between domains correspond to tree edges. Currently, COPS provides five classification layers called *Distant* (30% relative structural similarity), *Remote* (40%), *Related* (60%), *Similar* (80%), and *Equivalent* (99%) (1, 9).

The graphical interface requires JavaScript to be enabled as well as a recent (version 10 or greater) Adobe® FlashPlayer® installation. For the proper three-dimensional (3D) visualization of protein structures and superimpositions, we recommend a modern workstation with a minimum display resolution of 1,024×768 pixels and a fast network connection. COPS is available online at http://cops.services.came.sbg.ac.at/.

At start up the first COPS page shows a widget where the main tools such as qCOPS, iCOPS, and ΔCOPS are listed. This tutorial is focused on the first application, quantitative COPS (qCOPS). A typical COPS query involves several steps (refer to Fig. 1 for a condensed view):

1. Main Query
   Enter a PDB four letter code (e.g., 2hhb) into the query input box (Fig. 2a) and press the button *Search* or the return/enter key on your keyboard. This queries the qCOPS server with the given PDB code. In this tutorial, we use 1z6t (10) as our query.
2. Selection Widget (Fig. 2b)
   The result of a query is listed in the *Selection Widget* which displays all COPS domains available for a given PDB code.

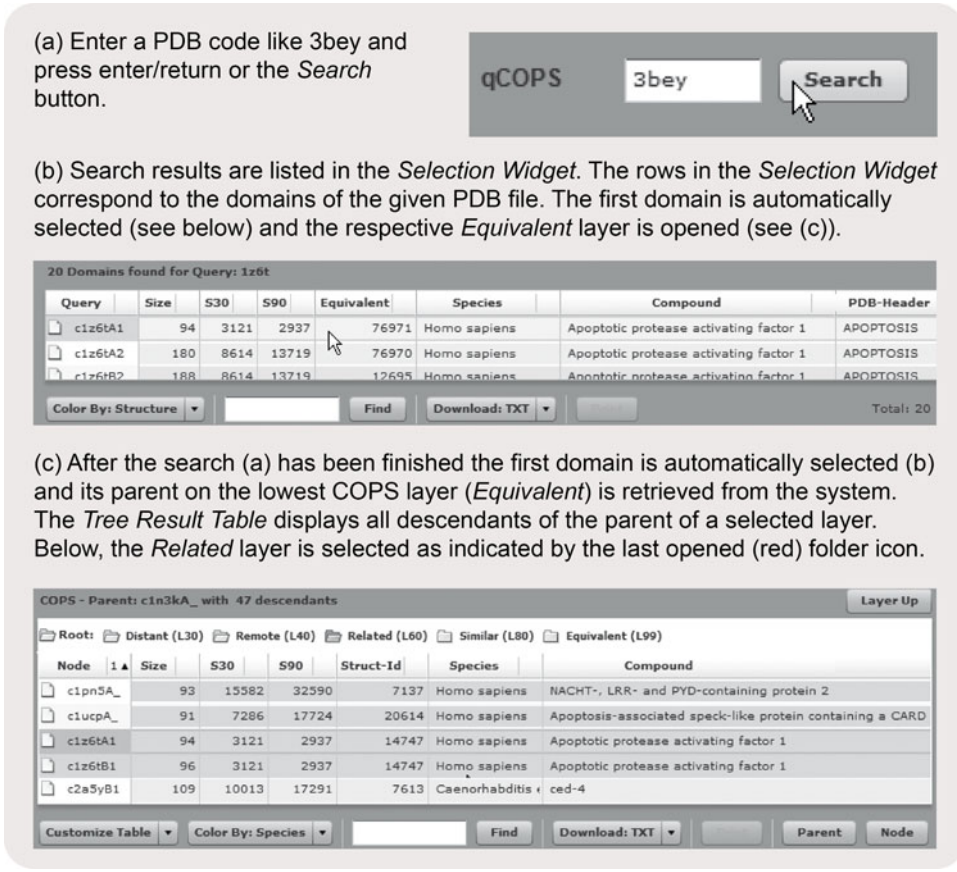

Fig. 1. The essential steps to use COPS.

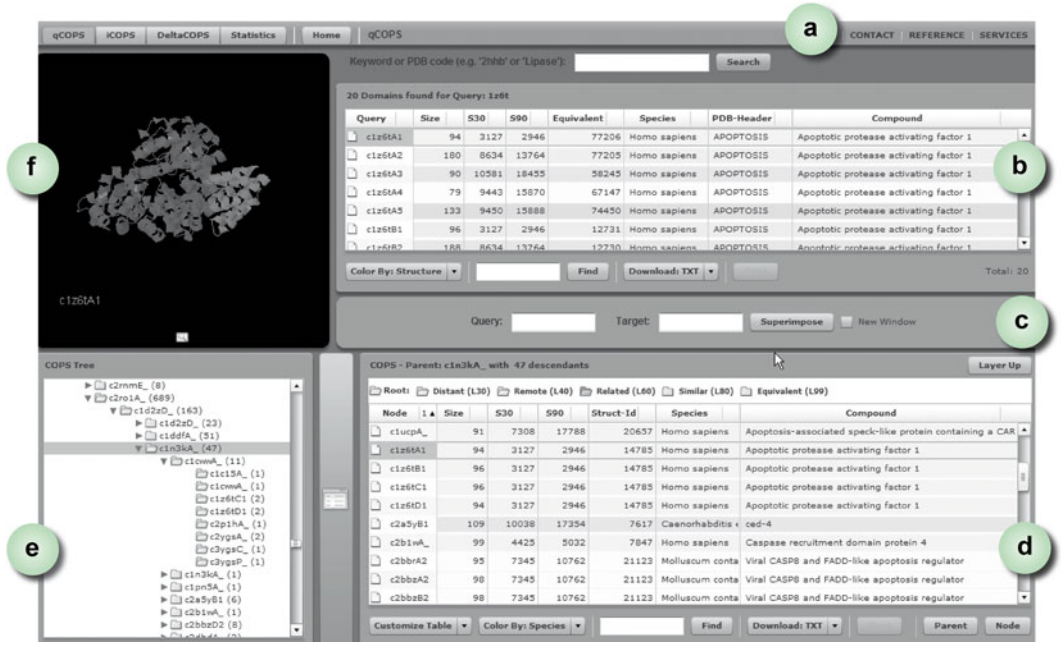

Fig. 2. COPS screen capture displaying the main sections of the interface: (**a**) Query input box, (**b**) Selection Widget, (**c**) Superimposition Box, (**d**) Tree Result Table, (**e**) Tree Widget, and (**f**) Jmol Widget.

Table 1
Table columns available in the *Selection Widget*[a] and the *Tree Result Table*[b]

Column	Description
Query/Node[a,b]	Unique domain name (see text for details)
Size[a,b]	Size of the domain in residues
S30[a,b]	Sequence classification code on layer S30. Domains with the same *S30* id are in the same sequence cluster and share at least 30% sequence identity
S90[a,b]	Sequence classification code on layer S90. Same as *S30*, but sequences within the same cluster share at least 90% sequence identity
Equivalent[a]	Structure classification code on the Equivalent layer (L90)
Struct-Id[b]	Structure classification code on the subsequent layer
Species[a,b]	Scientific name of the source organism used by UniProt and NCBI
PDB-Header[a,b]	HEADER classification record of the respective PDB file
Compound[a,b]	Describes the macromolecular contents of an entry
Method[b]	Experimental method
Resolution[b]	Resolution in Å
SG[b]	1 for Structural Genomics target, 0 otherwise
S-Kingdom[b]	Super Kingdom as defined in the NCBI taxonomy
Ligand Short[b]	Ligand short name
Ligand Long[b]	Ligand name
EC Number[b]	Enzyme classification number
Release Date[b]	Release date of the respective PDB file

Two actions are triggered as soon as the data of the *Selection Widget* has been loaded: First, the first domain is selected and visualized in context with the respective protein chain in the *Jmol Widget* (Fig. 2f), and second, the first domain is selected on the equivalent layer in the *Tree Result Table* (Fig. 2d) of the *Fold Space Navigator* (see below).

(a) The *Selection Widget* has a title bar where the query code and the number of domains are indicated. Every domain in the *Selection Widget* is annotated as described in Table 1. Domains are identified by a unique name constructed as follows: The first character is *c* followed by the four letter PDB code. The next letter specifies the PDB chain and the last letter numbers the domains within the chain. Single chain domains have an underscore as last character. For example, the code c1z6tB2 specifies domain two of chain B of PDB code 1z6t. Domains can be selected by clicking on the corresponding row in the table.

(b) The table rows are sorted by the domain names (*Query* column) by default. To sort the rows by any of the other columns just click on the respective column header. This is indicated by a small black triangle besides the column name which is visible when the column is sorted and the mouse pointer is placed over a column header. If the triangle points up the table is sorted in ascending order, if the triangle points down the sort order is descending. Additionally, a number is placed besides the triangle. This number indicates the sort order of the columns. For example, if the table rows are sorted by the *S30* column, a black triangle is visible in the *S30* column header together with the number one besides the column name. The number one indicates that column *S30* is the first sort criterion. We can now sort the table by a second criterion, e.g., the *Equivalent* column. This can be achieved by placing the mouse over the *Equivalent* column header and clicking on the number two appearing on the right side of the column name. Now the table rows are sorted or grouped firstly by the *S30* id and secondly by the *Equivalent* id. In other words, domains with more than 30% sequence identity are grouped together and these groups are then divided into subgroups of domains with more than 99% structural similarity. Other columns can be added to the sort criteria in the same fashion. To reset the sort criteria to the default sort order, just click on the column header of the *Query* column. More examples of useful sort combinations are given in the *Tree Result Table* paragraph of item 3.

You can also change the order of the columns in the table by dragging the column at the column header and dropping it at the desired position. To change a column width, place the mouse pointer over the grid lines separating two column headers and move the line with the appearing new mouse cursor to the desired width.

(c) Below the *Selection Widget* a toolbar is located that allows some customizations of the table. It is separated into three sections by pale vertical lines. With the drop-down list in the first section the table can be colored by different criteria. By default, the table is colored by *Structure*, which means all domains that share the same classification id on the *Equivalent* layer have the same color. In other words, domains in the same *Equivalent* layer are colored similarly. All columns (except *Query*) can be used for coloring the table. The coloring gives a quick overview of the domain composition of a protein and helps answering questions on the structural diversity of the domains. If we sort the domains of our example protein 1z6t by the *Equivalent* column and color by *Structure*, we instantly see that domains three, four, and five of chains A–D are structurally equivalent.

The next section of the toolbar is for searching the table with a domain name. For example, to get the third domain of chain C of 1z6t one can enter c1z6tC3 and click the *Search* button. The last section of the toolbar provides the data of the result table in different file formats such as CSV or XML.

3. Fold Space Navigator
   The *Fold Space Navigator* is a graphical representation of qCOPS and its design is largely equivalent to the structure of a file browser. Folder icons represent parent nodes (representative domain) on a given layer and the contents of a folder (i.e., the files) correspond to all child nodes (i.e., the complete subtree) of the respective family. The *Tree Widget* displays the path of the selected domain from the root (no structural similarities) of the hierarchical classification tree down to the equivalent layer (highest structural similarities). The structural relationship of all child nodes to the parent depends on the selected layer. On the equivalent layer, for example, all domains of a specific family have a structural similarity of ≥99% to the parent. The *Fold Space Navigator* contains three widgets: The *Tree widget*, the *Tree Result Table*, and the *Breadcrumb* for easy layer navigation. In the following, all three widgets are explained in detail.
   (a) Tree widget (Fig. 2e)
   The *Tree Widget* is hidden by default to maximize the *Tree Result Table* view. To uncover the *Tree Widget* just press the button on the left side of the *Tree Result Table*. The *Tree Widget* provides direct access to the nodes of the qCOPS hierarchy. Every icon folder corresponds to the parent domain on a specific layer. Besides an icon folder, the domain name of the representative domain (parent) is shown followed by the total number of child domains below the respective parent in parenthesis. Clicking on a folder icon loads the child domains into the *Tree Result Table*. The black arrows in front of the folder icons can be used to open or close a folder without loading the child nodes. Folder icons can be dragged and dropped into the *Superimposition Box* to get a structure alignment as we will see later (see item 4).
   (b) Tree Result Table (Fig. 2d)
   The *Tree Result Table* lists all child domains of a selected parent. The name of the parent and the number of descendants are displayed in the title bar of the table. The functionality of the table is similar to the result table of the *Selection Widget* (*see* item 2), but covers more columns and additional features. By default, the displayed columns are identical, except for the *Node* and the *Struct-Id* column. The *Node* column comprises domain names, too, but here it specifies the node names in the context of the classification tree. The *Struct-Id* column contains the layer id of a node on the subsequent layer (from root to leaf) or, if the

current layer is the *Equivalent* layer, the id of the (leaf) node itself. As a consequence, nodes on the *Equivalent* layer have all unique *Struct-Id* values. The representative domain (parent) of the currently selected layer has a folder icon besides the *Node* name that distinguishes it from the other domains in the table. Clicking on a row in the *Tree Result Table* displays the TopMatch superimposition of the respective node and the selected domain in the *Selection Widget* and the *Jmol Widget*.

Using the sort combinations explained in item 2, it is easy to answer difficult questions with just a few clicks. For example, suppose we are interested in domains that have relative structural similarities of at least 60% but sequence identities below 30%. We use domain one (c1z6tA1) of chain A of our example structure 1z6t. We skip the *Equivalent* and *Similar* layers and directly select the *Related* layer in the *Breadcrumb* navigation (see item 3c). Sort the table by the *Struct-Id* column by clicking on the respective column header and add the *S30* column as the second sort criterion as explained in item 2. Now we only have to scroll through the table and search for domains with identical *Struct-Id* but different *S30* entries. This process can be simplified even more by additionally coloring the table by *Structure*; then we only have to search for table rows with identical color but different *S30* values. In our example, numerous pairs of domains fulfill these criteria. To check the results, e.g., c3lqrA1 and c2vgqA4, we simply superimpose the domains with TopMatch (see item 4). In fact, the domains have almost 80% relative structural similarity but less than 15% sequence identity.

The *Tree Result Table* has a toolbar, similar to the toolbar of the *Selection Widget* (item 2). The functionality is identical except for the *Customize Table* button. This button opens a menu that enables the user to add or remove columns from the *Tree Result Table* by checking or unchecking the corresponding check boxes, respectively (see Table 1 for a column description). The buttons *Parent* and *Node* at the right end side of the toolbar select the parent and the node row (the currently selected domain in the *Selection Widget*) in the *Tree Result Table*.

(c) Breadcrumb Navigation (Fig. 2d)
The *Breadcrumb Navigation* widget above the *Tree Result Table* displays the path of the selected domain from the root (no structural similarities) of the hierarchical classification tree down to the equivalent layer (highest structural similarities). Each node of a layer on the path is depicted as a folder icon (cf. *Tree Widget*) followed by the layer name and the layer shortcut in parenthesis. The currently selected layer is highlighted red. A click on one of the folder icons

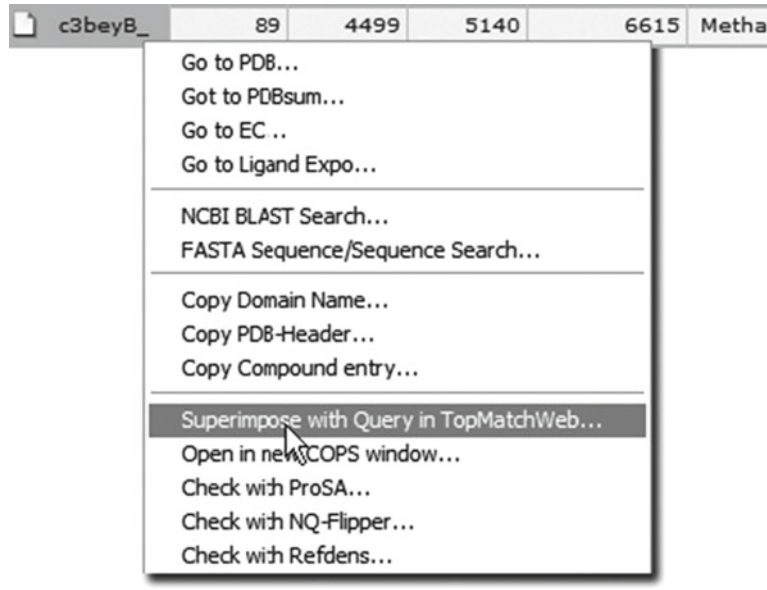

Fig. 3. The right-click context menu of the *Tree Result Table* is split into four sections. The first section contains entry-specific links to external resources such as PDB, PDBsum, Enzyme Classification (EC), Ligand Expo, and Pubmed (Primary Citation). The second section provides sequence search functionality and sequence data. Copy functionality is given in the third section, and the last section includes links to resources for structure comparison, structure search, and structure validation. For example, the first entry in the last section opens up a new window with the TopMatch (8) superimposition of the query and the selected target from the *Tree Result Table*. The second entry in the last section (*Open in new COPS window* …) queries COPS with the selected target from the *Tree Result Table* in a new window.

selects the representative domain on the respective layer and all descendants of the representative are listed in the *Tree Result Table*. The name of the parent is shown within the tool tip that appears when the mouse pointer is placed over the respective layer icon. It is identical to the entry with the folder icon in the *Tree Result Table* (item 3b). The *Breadcrumb Navigation* is automatically updated if the selection in the *Tree Widget* or the *Selection Widget* is changed.

4. Superimposition Box (Fig. 2c)
   The *Superimposition Box* provides access to the TopMatch structure alignment server (8). *Query* and *Target* name for the structure alignment have to be provided in the correspondingly named text fields. Domain names can be entered directly into the text fields or, more conveniently, dragged and dropped into the respective text fields. Drag and drop is possible from any widget with domain names, particularly the *Selection Widget*, the *Tree Widget*, and the *Tree Result Table*. Once the *Query* and *Target* fields are filled in, a click on the *Superimpose*

button opens a new browser window where the detailed TopMatch structure alignment is displayed. The TopMatch superimpositions are always loaded into the same external window as long as the *New Window* check box besides the button is not selected.

5. Jmol Widget (Fig. 2f)
   The *Jmol Widget* contains Jmol (http://www.jmol.org/), an open-source Java viewer for chemical structures in 3D. Below the applet a small magnifier is located that can be used to maximize the 3D view. Additionally, the maximized view displays the ligands of the respective chain, too.

3. Application of COPS in Homology Modeling

The major goal in homology modeling is to obtain an accurate structural model for a given protein sequence with unknown structure. The first step on the way to the model is the identification of proper structural templates for the given sequence. This is an essential step, since the template structures form the basic framework upon which the model is constructed. Hence, the choice of the templates has a significant impact on the quality of the resulting model.

The first step in homology modeling is the identification of evolutionary-related proteins with known structure that can serve as suitable templates for a specific target sequence. There is a plethora of sequence-based homology detection methods available for this task (11) with distinct capabilities in detecting homologous sequences (12). In general, all methods return a hit list sorted by a similarity score indicating the relevance of the specific hits. Hits within a certain threshold are considered to be trustable results and those with available structure files are potential templates for protein core modeling.

Table 2 shows the hit list for CASP8 target T0408 (http://predictioncenter.org/casp8/target.cgi?id=23&view=all) obtained by the sequence-based HHsearch algorithm in a search against a nonredundant template data base (13). Recently, HHsearch outperformed other sequence-based algorithms in an analysis of sequence database search methods (12). Entries from the hit list within the trustable cutoff (Table 2) are our potential templates in the modeling process of T0408. At this point of the modeling procedure, nothing is known about the structural similarities between the template candidates, their domain organization and other structural characteristics that facilitate the selection of templates for subsequent model building.

In the process of homology modeling, COPS can be applied as soon as the first template candidates have been identified. These structures can then be analyzed in terms of structural relationships

Table 2
HHsearch results for CASP target T0408 retrieved from the HHsearch web server (13) using default parameters

No	Hit		Prob	*E* value	SeqId (%)
1	3d7i_A	Carboxymuconolactone de	100.0	7.2E–32	97
2	3bey_A	Conserved protein O2701	100.0	2.2E–28	20
3	1p8c_A	Conserved hypothetical	99.9	1.8E–24	19
4	2qeu_A	Putative carboxymuconol	99.9	3.1E–24	23
5	2af7_A	Gamma-carboxymuconolact	99.9	1E–24	20
6	1vke_A	Carboxymuconolactone de	99.9	2.6E–24	18
7	2cwq_A	Hypothetical protein TT	99.9	2E–22	23
8	2q0t_A	Putative gamma-carboxym	99.9	1.6E–21	20
9	2q0t_A	Putative gamma-carboxym	99.9	3.4E–21	21
10	2ouw_A	Alkylhydroperoxidase AH	99.7	3.1E–16	22
11	1gu9_A	Alkylhydroperoxidase D;	99.7	2.5E–16	13
12	3c1l_A	Putative antioxidant de	99.3	1.1E–10	10
13	2prr_A	Alkylhydroperoxidase AH	99.2	2.3E–10	13
14	2gmy_A	Hypothetical protein AT	99.2	1.2E–10	15
15	2o4d_A	Hypothetical protein PA	99.2	2E–10	14
16	3lvy_A	Carboxymuconolactone de	99.0	1E–09	8
17	2pfx_A	Uncharacterized peroxid	99.0	1.9E–09	6
18	2oyo_A	Uncharacterized peroxid	99.0	2.9E–09	9
19	1gu9_A	Alkylhydroperoxidase D	97.9	0.00015	12
20	3bjx_A	Halocarboxylic acid deh	97.6	5E–06	14
21	2pfx_A	Uncharacterized peroxid	96.7	0.003	15
22	3lvy_A	Carboxymuconolactone de	96.1	0.0088	21
23	2oyo_A	Uncharacterized peroxid	96.1	0.004	14
24	2gmy_A	Hypothetical protein AT	95.9	0.0095	8
25	2o4d_A	Hypothetical protein PA	95.9	0.0063	16

The hit list is sorted by the estimated probability (Prob) which is the most important criterion for homology. According to the HHsearch manual hits with a probability larger than 95% are nearly certainly homologous to the query sequence. Therefore, only hits above the 95% probability cutoff are included. Additionally, the *E* value and the sequence identity (SeqId) to the query sequence are shown. The structure of T0408 has been solved by X-ray crystallography and is available as PDB file 3d7i.

to other proteins in the PDB, as well as structural differences between the templates (see Subheading 3.1). Furthermore, the candidates can be characterized by features describing their biological context, like source organism or functional annotation (see Subheading 3.2). We exemplify the practical usage of COPS for homology modeling in the following two subsections using the templates from Table 2 and other examples.

3.1. How Diverse Are My Template Structures?

The protein structures in Table 2 are putative templates for our model. Hits with the highest score and *E* value are considered to be the best templates. However, nontrivial templates (query coverage ≤ 90% and sequence identity ≤ 90%) may have structural varieties that are not detectable from the initial template list, but that are essential for model building. Structure comparison of the templates is an indispensable step in the process of template selection and alignment correction. This is especially useful if the structural differences are visualized and the corresponding sequence alignments are available. Pairwise structural comparisons and their visualizations are cumbersome tasks, but COPS and TopMatch facilitate this process considerably.

The first hit in the template list (Table 2) is the solved structure of target T0408 as determined by X-ray crystallography and deposited in the PDB with the code 3d7i (14). Since this structure was not available during prediction season in CASP8, we perform a COPS search with the second hit, 3bey (15). After the search has been finished, all six structural domains of 3bey are listed in the *Selection Widget* (Fig. 2b), the first domain in the list (c3beyA) is selected and visualized in the *Jmol Widget*, and all domains of the respective *Equivalent* layer are displayed in the *Tree Result Table*. It is obvious from the COPS domain names that all six domains of 3bey are single chain domains, because no domain numbers are given but underscores. The found domains have at least 90% sequence identity indicated by identical *S30* and *S90* values. If we stain the domains by the *Structure* column entries it is easy to see that the domains are in different *Equivalent* layers except for c3beyC_ and c3beyF_, thus their relative structural similarities are less than 99%. The data from the *Selection Widget* addresses the internal organization and domain composition of a given protein structure. The data from the *Tree Result Table* explained in the following paragraphs deals with the structural similarities to other domains in the protein space.

The main goal of this section is to investigate the structural differences and similarities between our template candidates. Templates that cover the same regions of the target sequence are descendants of the same parent domain and can be found in the same layers of the *Tree Result Table*, presumed that they share the same structure. In this case, it is most straightforward to start with

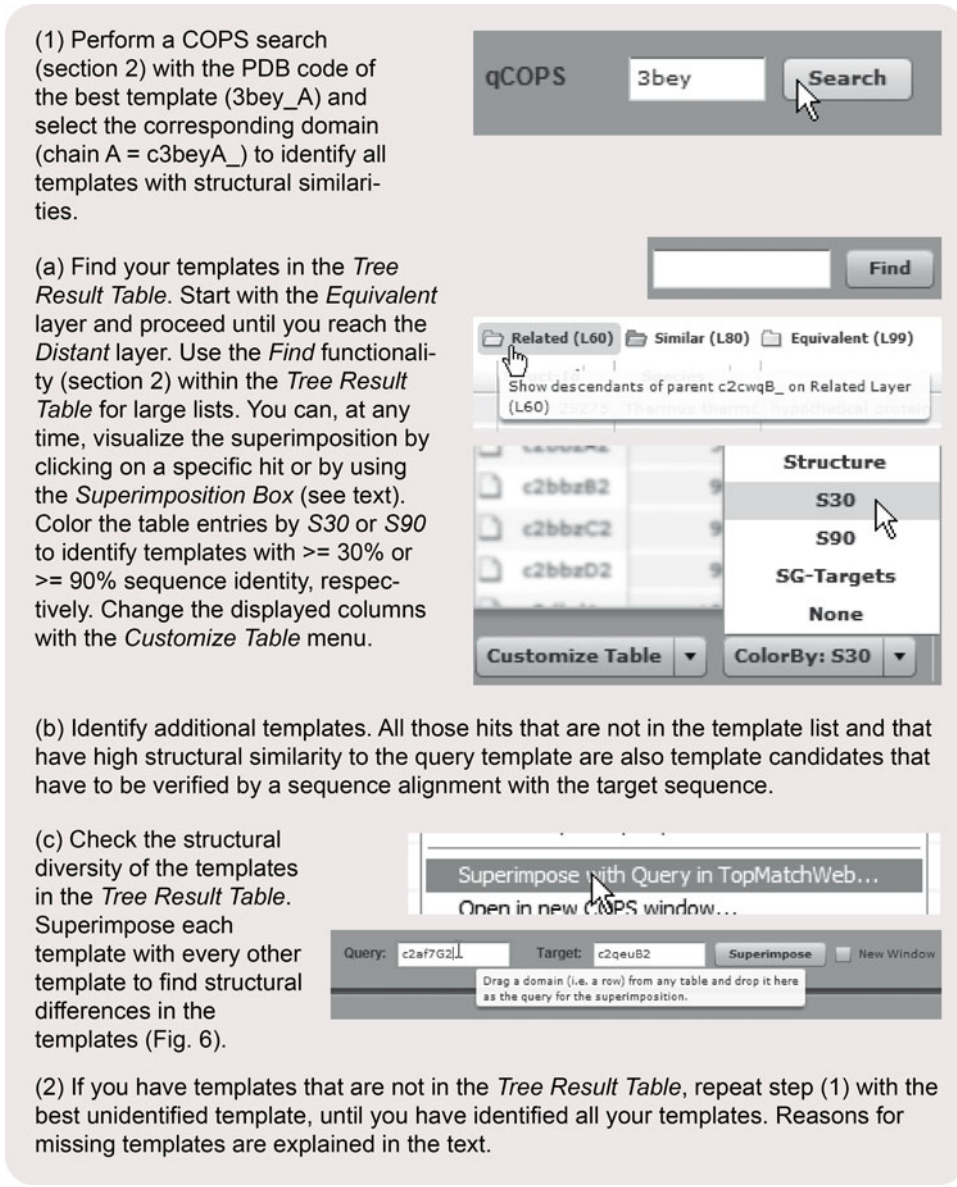

Fig. 4. Basic steps to investigate the structural diversity of a set of modeling templates. For details on the example used here, see Subheading 3.

the first template, browse through the hierarchical layers in COPS and identify the template structures from our template list from Table 2 For a condensed how-to manual of the following steps, refer to the box in Fig. 4.

The *Equivalent* layer of c3beyA_ contains one member and that is the domain itself. We switch to the next higher layer, the *Similar* layer, by clicking on the respective folder icon in the *Breadcrumb Navigation*. The parent c2cwqB_ on this *Similar* layer

has nine descendants including itself. Six domains are from 3bey (i.e., chains A–F) and three domains are from PDB file 2cwq (i.e., chains A–C) (16). If we color the *Tree Result Table* by *S30*, we see that the domains of 3bey and 2cwq are in different *S30* sequence clusters that means the domains have less than 30% sequence identity. As a consequence, the domains of the two PDB files are in different *S90* clusters, too.

All three chains (A–C) of 2cwq are stored as single chain domains within COPS. More than 90% of the domain sequences are identical illustrated by equivalent *S90* ids. In the template list, 2cwq is represented by template seven (i.e., chain A or c2cwqA_ in COPS, respectively). Generally, not all domains (respectively chains) from the *Tree Result Table* have to be comprised in the template list, since similar templates are pooled by HHsearch. Within the *Tree Result Table*, it is straightforward to validate the pools by checking the sequence and structure layers. Moreover, additional data is available to select the appropriate template from a pool. Columns that contain essential information supporting template selection and validation include experimental method, resolution, and the ligand columns. We will cover specific COPS columns in more detail where applicable.

A mouse click on the row of c2cwqA_ in *Tree Result Table* displays the TopMatch superimposition of the two templates c2cwqA_ and c3beyA_ (in COPS called target and query, respectively) in the *Jmol Widget*. The visualization of the superimposition and the respective layer give a first clue about the structural differences and similarities between the two templates (see Fig. 5c). For a detailed investigation, it is advisable to switch to the TopMatch server using the *Superimposition Box* (see Subheading 2, item 4 for details). Instantly, the same TopMatch superimposition is opened in an additional browser window, together with the structure-based sequence alignment and all key values of the alignment. In the structure-based sequence alignment, the structurally equivalent regions are colored red and orange, respectively, and the conserved residues are accentuated with black vertical bars. The 3D position of any amino acid in the protein structure can be highlighted by moving the mouse over the corresponding entry in the alignment. Together with the visualization of the ligands, these structural alignments greatly assist the identification of the structural core of the templates, as well as the validation of multiple sequence alignments of the templates.

To identify more templates in the *Tree Result Table*, we switch to the next higher layer, the *Related* layer. The parent domain remains the same (c2cwqB_), but the number of descendants increases to 36, because the structural similarity cutoff on the *Related* layer shrinks to 60%. We use the *Find* button to identify remaining templates. In addition to the already identified template c2cwqA_ from the *Similar* layer, templates three to six (1p8c_A,

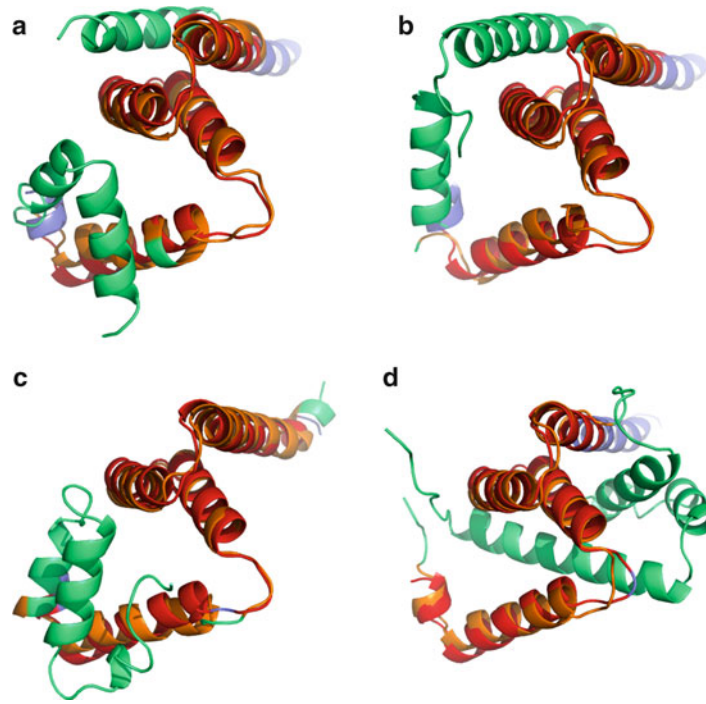

Fig. 5. Structural diversity among templates for CASP8 target T0408. The best hit (c3beyA_) from the HHsearch template list is superimposed with (**a**) c2af7A_, (**b**) c1vkeA_, (**c**) c2cwqA_, and (**d**) c2gmyA_. The first structure (query, here c3beyA_) is shown in *blue*, the second structure (target) in *green*, and the regions of similar structure are colored *red* (query) and orange (target).

2qeu_A, 2af7_A, and 1vke_A) are now present in the *Tree Result Table* of the *Related* layer. Again, we click on the rows of the respective templates to visually investigate the structural differences between the query (c3beyA_) and the other templates in the *Tree Result Table*. For example, structure 1p8c_A (17) is the second best template from the HHsearch template list (Table 2). Selecting the row of c1p8cA_ in the *Tree Result Table* displays the TopMatch superimposition of c1p8cA_ on c3beyA_. The superimposition in Fig. 6a reveals the structural similarity of c1p8cA_ and c3beyA_. c1p8cA_ covers 82% of c3beyA_ with an RMS of 1.8 Å, although the respective sequences have only 30% identical residues. Major structural differences are located at the carboxyl terminus (C terminus), where about half of the C-terminal α-helix of c3beyA_ is not superimposeable with c1p8cA_. This is the consequence of an almost 180° collapse in the α-helix of c1p8cA_, whereas the α-helix of c3beyA_ is elongated (see Fig. 6a). These unaligned regions are colored blue and green in the TopMatch alignment (Fig. 6a, b). One can easily determine the borders of the not superimposeable α-helices from the 3D view by moving the mouse over the sequences in the alignment. Here we have to decide if c1p8cA_ or c3beyA_ is

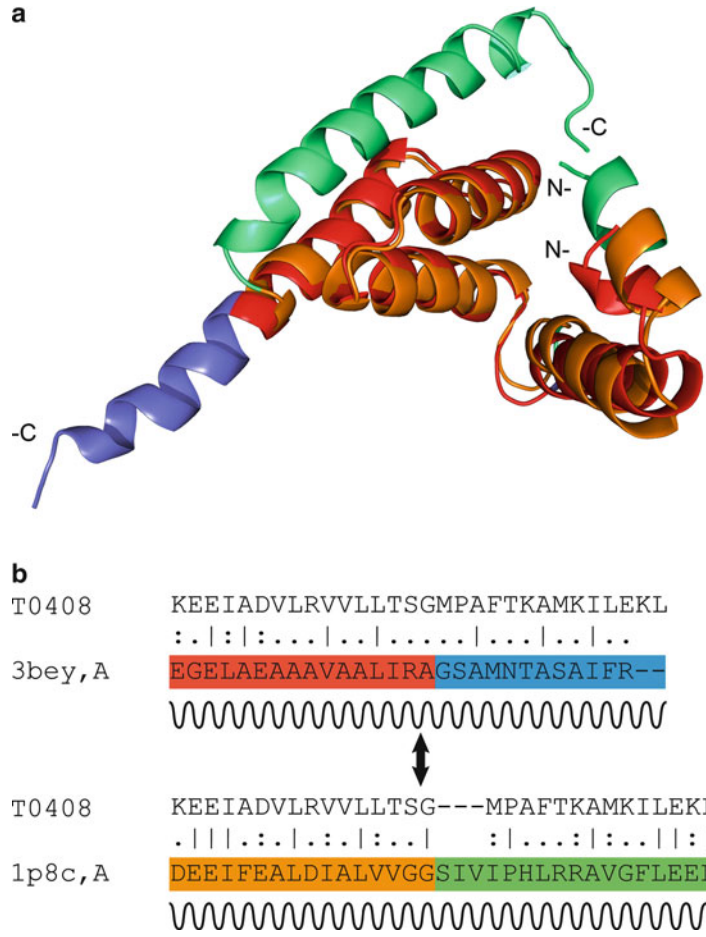

Fig. 6. Structural differences between the two best HHsearch templates for CASP target T0408 (Table 2). (**a**) TopMatch superimposition of first template 3bey,A (*blue* and *red*) with second template 1p8c,A (*green* and *orange*). *Red* and *orange parts* are structurally equivalent. The long C-terminal α-helix of 3bey,A cannot be superimposed on the corresponding α-helix of 1p8c,A over the full length of the helix. The reason is a considerable twist at residue GLY92 in 1p8c,A that involves an almost 180° collapse in the helix. (**b**) Pairwise sequence alignments of the C-terminal α-helices of the two templates with the target sequence (T0408). The *color coding* matches the TopMatch coloring from (**a**). The *black arrow* denotes the helix collapse. *Vertical bars mark* identical and *double dots* similar residues. Pairwise alignments were generated with EMBOSS (18).

the better template or if both structures are inadequate templates for this region. Best practice is to generate a pairwise sequence alignment of both templates with our target sequence (use the right-click menu explained in Fig. 3 to retrieve a specific protein sequence). Then the earlier defined borders of the respective α-helices from TopMatch can be identified in the pairwise sequence alignments (Fig. 6b). The target-template alignment shows higher sequence similarity at the collapsed α-helix of c1p8cA_ than at the

elongated α-helix of c3beyA_. To play it safe, one would use both templates to generate different models and examine the modeled structures with appropriate validation tools (c.f. Note 1).

It is highly advisable to proceed the whole template list in this fashion, at least for the best templates that are considered for modeling. In our case, the next template candidate is chain A of protein 2qeu (19). By repeating the previous steps, we are able to identify this entry as c2qeuA2 in the *Tree Result Table* in the same *Related* layer we discussed earlier. The domain name specifies c2qeuA2 as domain two of chain A of 2qeu. Obviously our query template 3bey,A has a different domain configuration as 2qeu,A, which can easily be verified by the TopMatch superimposition of the two domains. Three α-helices are perfectly superimposeable, but c2qeuA2 lacks the twist in the C-terminal α-helix (cf. c1p8cA_) and, additionally, the N-terminal α-helix of c3beyA_. The N-terminal α-helix is part of the first domain (c2qeuA1) of 2qeu,A. The same domain configuration can be found in the fifth best template 2af7_A. Both domains of 2af7 (c2af7A1 and c2af7A2) have highly similar structures compared to the two domains of 2qeu (relative structural similarity >80%), although c2qeuA2 and c2af7A2 are in different *S30* layers.

All templates from the template list can be found at least on the next higher layer, the *Remote* layer, except for the template 3bjx_A on position 20. Even on the *Distant* layer, which is the highest COPS layer beneath the *Root*, where the descendants have only 30% relative structural similarity to the parent, this protein structure is missing. In some cases, it is possible that templates from the template list cannot be found in the layers of the *Tree Result Table*; for instance if the templates are matching on different parts of the target sequence. In this case, it is advisable to use the first unidentified template in the COPS search, just like we used chain A of 3bey in the previous example. Moreover, this is indicative of templates that match different domains of the target sequence.

Another reason for missing templates in the *Tree Result Table* is structural diversity among the templates. In the worst case, the result is a false positive, like 3bjx,A from the template list. The sequence similarity scores returned for this template are all considered to be significant, but pairwise structural comparisons to the other templates reveal no trustable structural equivalences (see Fig. 7). A single template with no significant structural similarity to other templates in the list should be regarded with caution. If the sequence similarity to the target is weak, too, and the template covers the same regions of the sequence as other, more trustable templates, it is save to skip this structure.

Further reasons for missing templates in the *Tree Result Table* include protein structures with similar sequences but different 3D structures. We report more on this phenomenon in Note 2.

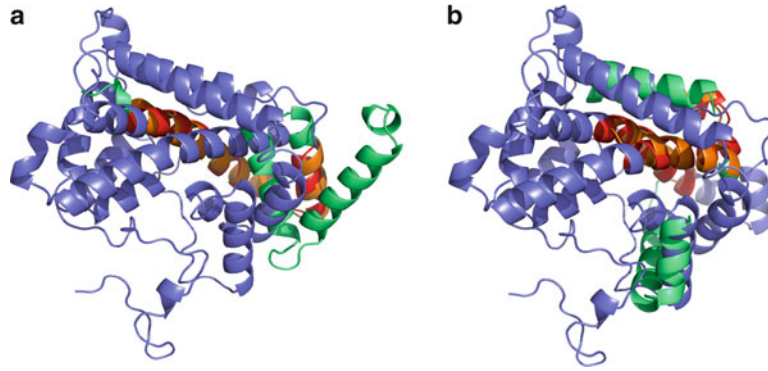

Fig. 7. Comparison of the potential template 3bjx_A (in *blue/red*) with (**a**) the best HHsearch template 3bey_A and (**b**) chain A of the released structure of CASP8 target T0408 (PDB code 3d7i). 3bjx_A is not a suitable template for T0408 although having significant scores (Table 2). More information about the characterization of potential false positives can be found in Subheading 3.1.

3.2. What Is the Biological Context of My Templates?

For many modeling targets, at least basic information is available about the biological context of the sequence, such as its source organism, its putative role in the cell or known binding partners. This information provides valuable clues for template selection in addition to sequence similarity and further data from experiments (e.g., chemical shifts, c.f. Note 3).

COPS domains shown in the *Selection Widget* or the *Tree Result Table* are annotated with several features that can be employed to narrow down the set of template candidates (see Fig. 8). For instance, the source organisms of the respective protein chains and their assignment to a taxonomic superkingdom can be compared across potential templates using the *Species* and *S-Kingdom* columns. Taking up our example above (T0408), we find that the target sequence was obtained from the archaeon *Methanocaldococcus jannaschii*. The HHsearch template list contains only two more proteins from archaea. The first is the highest ranking template 3bey_A and the second is structure 2af7_A at rank five; all other templates are from bacteria. In general, template structures from evolutionary-related organisms should be favored. Note, however, that a template from the same organism as the target sequence might have considerable changes in its fold, because proteins that result from the duplication of a gene (paralogs) are usually no longer subject to functional constraints (20–24).

The list of putative templates can also be characterized by functional aspects of the respective proteins. According to the *PDB-Header* column in COPS, the template list contains ten proteins with unknown function, eight oxidoreductases, and five lyases. Together with the more detailed *Compound* data this information can be used to find templates that match descriptions of function available for the target sequence.

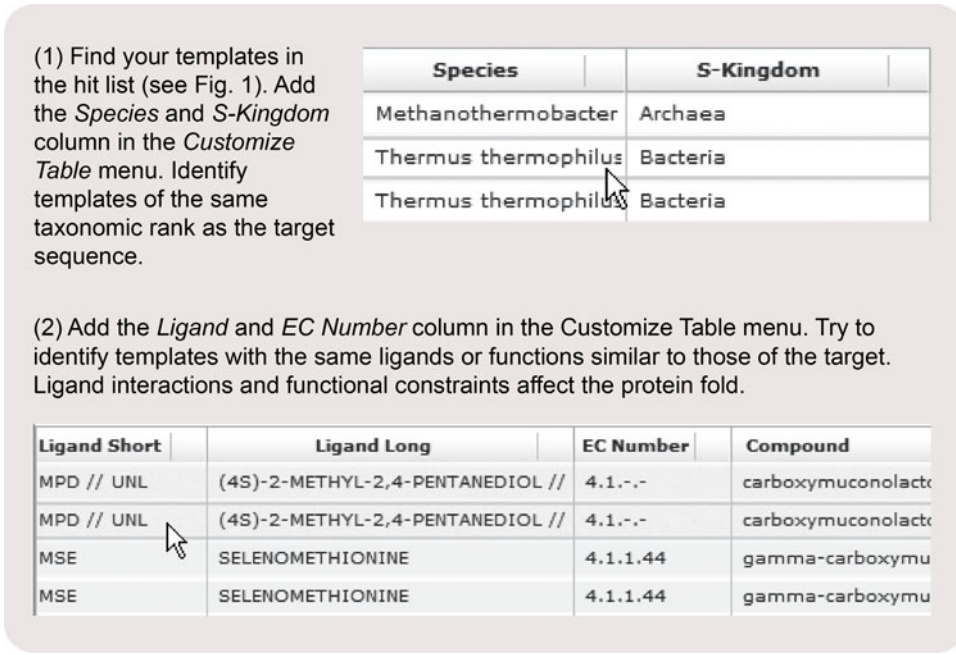

Fig. 8. Basic steps to investigate the biological context of putative template structures in COPS.

Ligands are another important source for clues on the biochemical function of proteins. They often affect the 3D structure of proteins resulting in considerable differences between the plain and the ligand bound conformations. Interfaces where ligands are bound depend on specific residues that interact with the ligand. Frequently, these residues are conserved across species. For example, the apoptotic protease-activating factor 1 (Apaf-1, PDB code 1z6t (10)) from *Homo sapiens* comprises five distinct domains in its chain A: (1) CARD, (2) an α/β fold, (3) helical domain I, (4) a winged-helix domain, and (5) helical domain II. Apaf-1 is bound to the ligand ADP. Three domains of Apaf-1 (the α/β fold, helical domain I, and the winged-helix domain) have equivalent domains in chain C of the apoptosis regulator *CED*-4-*CED*-9 (PDB code 2a5y (25)) from *Caenorhabditis elegans*. If superimposed pairwise, the equivalent domains have high structural similarities but sequence similarities below 30% (1). On chain level only the CARD domain and the α/β-fold can be superimposed simultaneously. This means that the arrangement of the domains in the protein chains is different for the ATP-bound 2a5y and the ADP-bound 1z6t. Both conformations are a consequence of the bound ligands. In particular, ADP locks Apaf-1 in the inactive conformation because it promotes the interactions between the domains of 1z6t (10). This is a clear example of how ligand binding can alter the structure of a protein. Even so, five residues of the eight residues that bind ADP and ATP, respectively, are conserved and structurally equivalent.

Regions of proteins that lack a well-defined three-dimensional structure may switch to an ordered state upon interaction with a

ligand (26). Automated methods may confusingly predict such regions as having a specific secondary structure as well as being disordered (27). If a template aligns to a region predicted to be disordered in the target, the ligand information given in COPS and the 3D visualization of their location in Jmol assist in the identification and validation of these regions.

To gather information on ligands in COPS and compare it across the templates, enable the *Ligand Short/Ligand Long* columns in the *Tree Result Table*. Additionally, the location of the ligands in the 3D structure can be visualized in the maximized *Jmol Widget* (Fig. 2f) and the external TopMatch window. The *Ligand* columns display all ligands associated with the respective PDB chain, separated by two slashes. In *Ligand Short*, ligands are represented by their shortcuts as defined by PDB. The entry *Go to Ligand Expo* in the context menu of the hit list links to the corresponding Ligand Expo page of PDB. This page offers 3D visualization of the selected ligand as well as detailed chemical and structural information. Enzymes in the *Tree Result Table* are further characterized by the entries in the *EC Number* column. This column contains the Enzyme Classification numbers as provided by the IUBMB (http://www.chem.qmul.ac.uk/iubmb/enzyme/). The detailed description of each enzymatic reaction can be opened with the *Go to EC* entry in the context menu of the *Tree Result Table*.

4. Notes

1. Final model quality is affected by a multitude of factors. Since each step in homology modeling implies its own pitfalls and error sources, it is vital to continuously check potential model structures for inaccuracies introduced by the modeling pipeline. In particular, care should be taken in template selection by choosing templates with high quality. Various parameters that can be used to winnow template structures in terms of quality directly originate from experimental structure determination, like crystallographic resolution or *R*-factor (28). In the *Tree Result Table* of COPS, the *Method* and *Resolution* columns can be consulted to get first clues on template quality. In addition, several tools directly linked from COPS provide independent quality estimates of potential template structures as well as the resulting models. ProSA (29, 30) employs knowledge-based potentials to recognize erroneous coordinates of protein structures. Besides a global quality measure, ProSA yields quality scores on residue level which allows to identify problematic parts of the template. Following a related approach, NQ-Flipper (31) recognizes unfavorable rotamers of asparagine and glutamine residues and provides means to download a corrected model. Side-chain correctness, in general, may be

analyzed by using a different approach (32) which compares local electron density distributions to their expected analogs. Using this method, it is possible to detect a wide variety of problems including unrealistic atomic contacts, unusual rotamers, and incorrect atom naming. Further computational tools widely used for model validation include Procheck (33), MolProbity (34), and WHAT_CHECK (35).

2. Currently only a few cases of pairs of proteins with high sequence similarity and different conformations are known, but this phenomenon may be more common than previously thought (36, 37). Designed proteins with these properties have been reported (38, 39), and there are also examples of naturally occurring proteins of this kind. Roessler et al. (40) found two members of the Cro repressor family having sequence identities as high as 40%, although half of their structures have switched from helices to strands. Moreover, some proteins have the ability to switch between several stable conformations (41–43). For instance, the chemokine lymphotactin adopts two distinct folds at equilibrium under physiological conditions (44). In the CASP6 experiment, the experimentally solved structure of one of the targets showed a conformation considerably different to that of the best template although having the same sequence (45). In a large-scale analysis with 13,000 protein chains (46), sequence alignment-based structural superpositions and geometry-based structural alignments for protein pairs were carried out to determine the extent to which sequence similarity ensures structural similarity. There were many examples where two proteins that are similar in sequence have structures that differ significantly. Some homology detection tools are searching against a nonredundant set of templates defined by sequence similarity. Important structure information for the modeling process can be lost if a nonredundant set of structures is constructed based merely on sequence similarity. TopMatch provides the possibility to perform both sequence-based superpositions and structure-based superpositions for a detailed investigation of such cases.
3. Chemical shifts are the “mileposts” of NMR spectroscopy (47). They are used for direct refinement of protein structures (48), prediction of protein secondary structure (49, 50), inference of protein backbone angles (51, 52), structure validation (53), and detection of structural similarities in proteins (54). Supplementing modeling by chemical shift information has gained interest (again) over the past years. In 2008, the CS23D Server (51) was presented which rapidly generates structures from both chemical shift and sequence information. In the beginning of 2009, Shen ea. (52) published a modified version of the structure prediction tool Rosetta which applies a chemical shift filter to improve the quality of the fragments used for

model generation. Finally, Ginzinger and Coles (55) published work on a fast structure database search which uses the chemical shifts of the target protein to reliably identify structural templates even in cases of low amino acid sequence similarity.

Acknowledgments

This work was supported by FWF Austria grant number P21294-B12.

References

1. Suhrer SJ, Wiederstein M, Gruber M, et al. (2009) COPS-a novel workbench for explorations in fold space. Nucleic Acids Res 37:W539–W544
2. Suhrer SJ, Wiederstein M, Sippl MJ (2007) QSCOP – SCOP quantified by structural relationships. Bioinformatics 23:513–514
3. Suhrer SJ, Gruber M, Sippl MJ (2007) QSCOP-BLAST–fast retrieval of quantified structural information for protein sequences of unknown structure. Nucleic Acids Res 35:W411–W415
4. Choi WS, Jeong BC, Joo YJ, et al. (2010) Structural basis for the recognition of N-end rule substrates by the UBR box of ubiquitin ligases. Nat Struct Mol Biol 17:1175–1181
5. Norambuena T, Melo F (2010) The Protein-DNA Interface database. BMC Bioinformatics 11:262
6. Berman HM, Westbrook J, Feng Z, et al. (2000) The Protein Data Bank. Nucleic Acids Res 28:235–242
7. Chothia C, Lesk AM (1986) The relation between the divergence of sequence and structure in proteins. EMBO J 5:823–826
8. Sippl MJ, Wiederstein M (2008) A note on difficult structure alignment problems. Bioinformatics 24:426–427
9. Sippl MJ, Suhrer SJ, Gruber M, et al. (2008) A discrete view on fold space. Bioinformatics 24:870–871
10. Riedl SJ, Li W, Chao Y, et al. (2005) Structure of the apoptotic protease-activating factor 1 bound to ADP. Nature 434:926–933
11. Cozzetto D, Kryshtafovych A, Fidelis K, et al. (2009) Evaluation of template-based models in CASP8 with standard measures. Proteins 77 Suppl 9:18–28
12. Frank K, Gruber M, Sippl MJ (2010) COPS Benchmark: interactive analysis of database search methods. Bioinformatics 26:574–575
13. Söding J (2005) Protein homology detection by HMM-HMM comparison. Bioinformatics 21:951–960
14. JCSG (2008) Crystal structure of carboxymuconolactone decarboxylase family protein possibly involved in oxygen detoxification (1591455) from *Methanococcus jannaschii* at 1.75Å resolution. To be published
15. Kuzin A, Xu JGX, Neely H, et al. (2007) Crystal structure of the protein O27018 from *Methanobacterium thermoautotrophicum*. To be published
16. Ito K, Arai R, Fusatomi E, et al. (2006) Crystal structure of the conserved protein TTHA0727 from Thermus thermophilus HB8 at 1.9 A resolution: A CMD family member distinct from carboxymuconolactone decarboxylase (CMD) and AhpD. Protein Sci 15:1187–1192
17. Kim Y, Joachimiak A, Brunzelle J, et al. (2003) Crystal Structure Analysis of *Thermotoga maritima* protein TM1620 (APC4843). To be Published
18. Rice P, Longden I, Bleasby A (2000) EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet 16:276–277
19. JCSG (2007) Crystal structure of Putative carboxymuconolactone decarboxylase (YP-555818.1) from *Burkholderia xenovorans* LB400 at 1.65Å resolution
20. Koonin EV (2005) Orthologs, paralogs, and evolutionary genomics. Annu Rev Genet 39:309–338
21. Pál C, Papp B, Lercher MJ (2006) An integrated view of protein evolution. Nat Rev Genet 7:337–348
22. Andreeva A, Murzin AG (2006) Evolution of protein fold in the presence of functional constraints. Curr Opin Struct Biol 16:399–408
23. Chothia C, Gough J (2009) Genomic and structural aspects of protein evolution. Biochem J 419:15–28

24. Worth CL, Gong S, Blundell TL (2009) Structural and functional constraints in the evolution of protein families. Nat Rev Mol Cell Biol 10:709–720
25. Yan N, Chai J, Lee ES, et al. (2005) Structure of the CED-4-CED-9 complex provides insights into programmed cell death in *Caenorhabditis elegans*. Nature 437:831–837
26. Dyson HJ, Wright PE (2005) Intrinsically unstructured proteins and their functions. Nat Rev Mol Cell Biol 6:197–208
27. Bordoli L, Kiefer F, Arnold K, et al. (2009) Protein structure homology modeling using SWISS-MODEL workspace. Nat Protoc 4:1–13
28. Wlodawer A, Minor W, Dauter Z, et al. (2008) Protein crystallography for non-crystallographers, or how to get the best (but not more) from published macromolecular structures. FEBS J 275:1–21
29. Sippl MJ (1993) Recognition of errors in three-dimensional structures of proteins. Proteins 17:355–362
30. Wiederstein M, Sippl MJ (2007) ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. Nucleic Acids Res 35:W407–W410
31. Weichenberger CX, Byzia P, Sippl MJ (2008) Visualization of unfavorable interactions in protein folds. Bioinformatics 24:1206–1207
32. Ginzinger SW, Weichenberger CX, Sippl MJ (2010) Detection of unrealistic molecular environments in protein structures based on expected electron densities. J Biomol NMR 47:33–40
33. Laskowski RA, MacArthur MW, Moss DS, et al. (1993) PROCHECK: a program to check the stereochemical quality of protein structures. J Appl Crystallogr 26:283–291
34. Chen VB, Arendall WB, Headd JJ, et al. (2010) MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr D Biol Crystallogr 66:12–21
35. Hooft RW, Vriend G, Sander C, et al. (1996) Errors in protein structures. Nature 381:272
36. Davidson AR (2008) A folding space odyssey. Proc Natl Acad Sci U S A 105:2759–2760
37. Sippl MJ (2009) Fold space unlimited. Curr Opin Struct Biol 19:312–320
38. Dalal S, Balasubramanian S, Regan L (1997) Protein alchemy: changing beta-sheet into alpha-helix. Nat Struct Biol 4:548–552
39. He Y, Chen Y, Alexander P, et al. (2008) NMR structures of two designed proteins with high sequence identity but different fold and function. Proc Natl Acad Sci U S A 105:14412–14417
40. Roessler CG, Hall BM, Anderson WJ, et al. (2008) Transitive homology-guided structural studies lead to discovery of Cro proteins with 40% sequence identity but different folds. Proc Natl Acad Sci U S A 105:2343–2348
41. Murzin AG (2008) Metamorphic Proteins. Science 320:1725–1726
42. Gambin Y, Schug A, Lemke EA, et al. (2009) Direct single-molecule observation of a protein living in two opposed native structures. Proc Natl Acad Sci U S A 106:10153–10158
43. Bryan PN, Orban J (2010) Proteins that switch folds. Curr Opin Struct Biol 20:482–488
44. Tuinstra RL, Peterson FC, Kutlesa S, et al. (2008) Interconversion between two unrelated protein folds in the lymphotactin native state. Proc Natl Acad Sci U S A 105:5057–5062
45. Ginalski K (2006) Comparative modeling for protein structure prediction. Curr Opin Struct Biol 16:172–177
46. Kosloff M, Kolodny R (2008) Sequence-similar, structure-dissimilar protein pairs in the PDB. Proteins 71:891–902
47. Zhang H, Neal S, Wishart DS (2003) RefDB: a database of uniformly referenced protein chemical shifts. J Biomol NMR 25:173–195
48. Schwieters CD, Kuszewski JJ, Tjandra N, et al. (2003) The Xplor-NIH NMR molecular structure determination package. J Magn Reson 160:65–73
49. Wishart DS, Sykes BD, Richards FM (1992) The chemical shift index: a fast and simple method for the assignment of protein secondary structure through NMR spectroscopy. Biochemistry 31:1647–1651
50. Wang Y, Jardetzky O (2002) Probability-based protein secondary structure identification using combined NMR chemical-shift data. Protein Sci 11:852–861
51. Berjanskii MV, Neal S, Wishart DS (2006) PREDITOR: a web server for predicting protein torsion angle restraints. Nucleic Acids Res 34:W63–W69
52. Shen Y, Delaglio F, Cornilescu G, et al. (2009) TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. J Biomol NMR 44:213–223
53. Oldfield E (1995) Chemical shifts and three-dimensional protein structures. J Biomol NMR 5:217–225
54. Ginzinger SW, Fischer J (2006) SimShift: identifying structural similarities from NMR chemical shifts. Bioinformatics 22:460–465
55. Ginzinger SW, Coles M (2009) SimShiftDB; local conformational restraints derived from chemical shift similarity searches on a large synthetic database. J Biomol NMR 43:179–185

http://www.springer.com/978-1-61779-587-9

Homology Modeling
Methods and Protocols
Orry, A.J.W.; Abagyan, R. (Eds.)
2012, XI, 419 p. 89 illus., 55 illus. in color., Hardcover
ISBN: 978-1-61779-587-9
A product of Humana Press