Chapter 2
Biomacromolecular Fragments and Patterns

Lukáš Pravda

The function of biomacromolecules such as proteins is intimately connected with their three-dimensional (3D) structure, and as such it is a reasonable starting point for structure-based drug design. Since the tertiary structure is more evolutionarily conserved than the primary sequence, the analysis of 3D structure provides key insights, not only in terms of classification, but has many implications in biotechnologies and drug design. On one hand, we can search for novel binding partners of characterized and validated target proteins; on the other hand, we can infer the function of as-yet uncharacterized proteins responsible for various diseases.

The question is, which part of a biomacromolecule or biomacromolecular properties do we want to evaluate? In general, we are mainly interested in the parts exhibiting biological functions. These are usually small and well-conserved spatial arrangements of amino acids and/or interacting ligands, such as cofactors; substrates or products of enzymatic reactions, inhibitors, or messenger molecules. In this book we collectively refer to these protein substructures as biomacromolecular patterns or fragments. A pattern can, in principle, take a number of different forms. It may be amino acids constituting catalytic or binding sites, sequence patterns responsible for cell signaling [1], allosteric regions [2–4], protein pockets and cavities [5–7], channel lining residues [8, 9] etc.

One of the first steps in every in silico analysis for not only drug design, is the detection of these biologically important patterns. There can be many reasons behind them. We can identify similar binding sites in off-target proteins,[1] discover new inhibitors, facilitate the identification of protein-protein interactions, or evaluate ligand-accessible pathways to the enzyme reaction site to name a few.

[1]Off-target protein binding implies an undesirable binding of a small molecule with a therapeutic effect to a protein target other than the primary target for which it was intended. Such binding often causes unintended side effects.

J. Koča et al., *Structural Bioinformatics Tools for Drug Design*,
SpringerBriefs in Biochemistry and Molecular Biology,
DOI 10.1007/978-3-319-47388-8_2

2.1 Pattern Examples

2.1.1 Active Site and Their Inhibition – Cyclooxygenase Inhibitors

The cyclooxygenase enzymes (COX-1 and COX-2) are responsible for the bis-oxygenation of arachidonic acid to prostaglandins. This process is critical during inflammation, cancer, but also in kidney development or maintaining gastrointestinal integrity and it is responsible for pain [10]. As such they are primary targets for a large body of nonsteroidal anti-inflammatory drugs such as aspirin or ibuprofen. While COX-1 is expressed constantly in the majority of cells and possesses a housekeeping function, COX-2 is only induced by inflammatory stimuli. Therefore, the development of selective inhibitors which can in turn be used for example as anti-inflammatory and anticancer agents with as few side-effects as possible, is of great interest. The first structures of COX enzymes were solved some 20 years ago, revealing the binding pocket and their inhibitors as highlighted in Fig. 2.1.

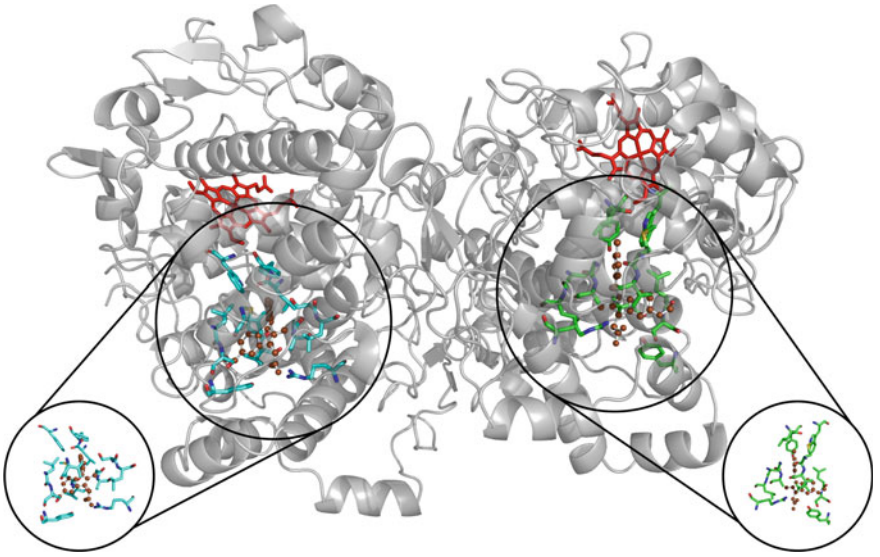

Fig. 2.1 Structure of COX-2 complexed with the indomethacin non-selective inhibitor (PDB ID `4cox`). COX-2 is a homodimer, with each unit containing a cyclooxygenase active site and a peroxidase active site. The peroxidase active site is involved in activating the heme group (*red*), which is crucial for further cyclooxygenase reaction. The molecular patterns of the COX-2 inhibitor (in *brown*) together with its interacting partners (*green* or *cyan* with respect to a protein unit) in the enzyme active sites are highlighted. The inhibitor is stabilized both by polar and nonpolar interactions (color figure online)

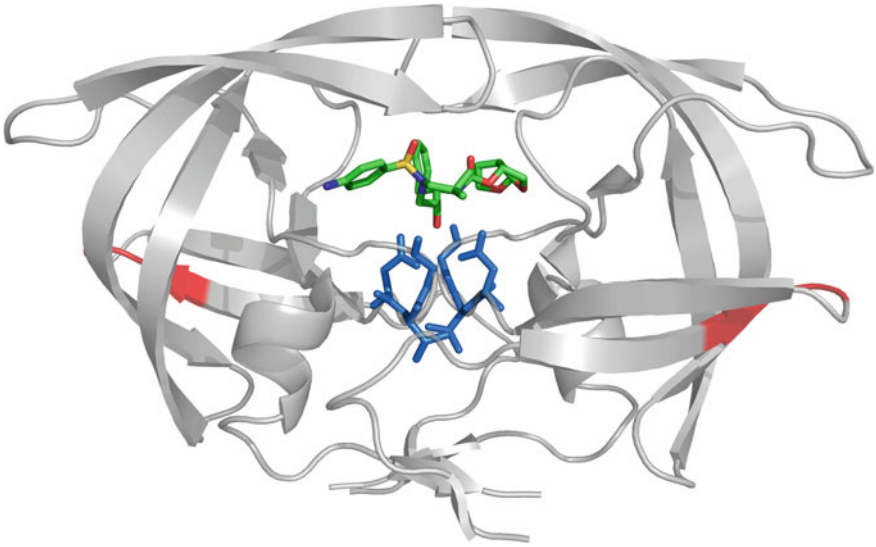

Fig. 2.2 HIV-1 protease complexed with the inhibitor darunavir (PDB ID `3lzv`). The molecular pattern of a catalytic triad is highlighted in *blue*. Elbow allosteric regions presumably responsible for the protein flexibility are shown in *red* (color figure online)

2.1.2 Allosteric Site – Structural Flexibility of HIV Protease

Inhibition of the HIV-1 protease is considered to be one of the three key avenues for blocking HIV replication, and therefore prevention of the development of AIDS [11]. Inhibition of the HIV-1 protease active site with drugs like ritonavir, nelfinavir, and amprenavir was considered to be an efficient approach. As a consequence of the drug binding, HIV-1 protease loses its dynamic behavior, which is crucial for its proteolytic function. However, many drug-resistant variants emerged, so inhibitor development continues. A number of NMR and MD experiments revealed putative regions responsible for the enzyme flexibility. As such these allosteric regions can be rationally targeted by novel allosteric inhibitors, in order to inactivate the enzyme's function [12]. Figure 2.2 displays a catalytic triad of the enzyme active sites together with putative allosteric sites responsible for the enzyme's flexibility.

2.1.3 Transcription Factor – Zinc Finger Motif

The DNA-binding class of enzymes called zinc fingers (ZnF) is the most abundant across all biota. The first classical ZnFs denoted as C_2H_2 were extracted from the *Xenopus* transcription factor, where they specifically bind DNA and control transcription [13]. Besides this, ZnFs are responsible for DNA recognition, the regulation of

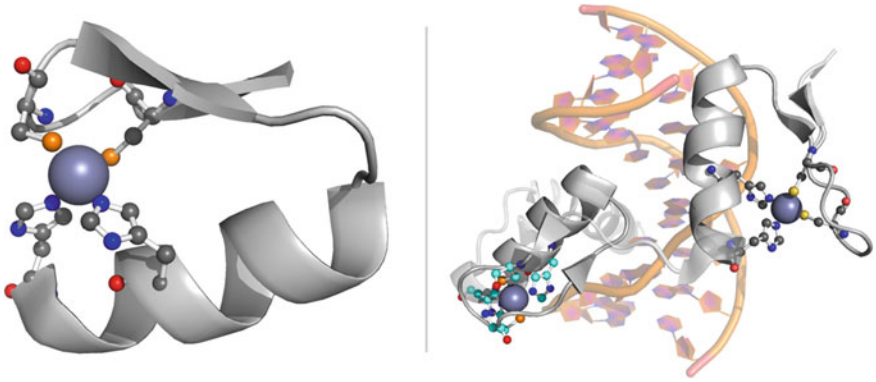

Fig. 2.3 C_2H_2 zinc finger motifs of the transcription factor early growth response protein 1 (Egr1) (PDB ID `4r2a`). The figure on the *left* depicts the overall cartoon model of a zinc finger motif, with the residues (two cysteines and two histidines) responsible for zinc ion binding. The zinc ion is shown as a sphere, while cysteine and histidine are denoted in a ball-and-stick model. In the other figure, two zinc fingers are bound to the major groove of the DNA strand

apoptosis and lipid binding. This motif is usually defined by a simple primary structure pattern called a consensus profile. Nevertheless, atypical motifs exist deviating from the consensus profile X_2-C-X_{2-4}-C-X_{12}-H-X_{3-5}-H (X stands for any amino acid, C is cysteine and H represents histidine in the consensus profile), that recognize specific genomic sites. The X_{12} region is usually further decomposed into the sequence X_3-[F|Y]-X_5-ψ-X_2, where [F|Y] represents either a phenylalanine or tyrosine residue, and ψ denotes a hydrophobic residue. At the 3D level, this sequence has a simple $\beta\beta\alpha$ fold, which is stabilized with a zinc ion coordinated with two histidine and two cysteine residues as shown in Fig. 2.3.

2.2 Pattern Prediction

Over the past few decades a plethora of software tools have been developed for the detection and extraction of biomacromolecular patterns from protein structures. The individual tools differ in the level of pattern description, the employed algorithms and of course their applicability. Drug design usually aims to identify potential binding sites in target and off-target proteins. These are often located in shallow protrusions in the protein surface referred to as pockets or clefts, as well as deeply buried in the protein structure. Therefore, the majority of the software is designed for predicting suitable pockets in apoproteins and holoproteins (e.g. CASTp [14], Pass [15], Q-SiteFinder [16], or FTSite [17]). Others may identify accessible pathways for the small ligands interacting with the proteins (e.g. MOLE 2.0 [18], Caver 3.0 [19] or MolAxis [20]). These are discussed in more detail in Chap. 6 – Detection of

Channels. Generally, pocket prediction for binding protein inhibitors can be classified into two groups: *geometry-based algorithms* and *energy-based algorithms*.

The geometry-based algorithms involve a couple of approaches. The most popular group of algorithms involves the projection of the protein structure onto a *3D grid* with a custom spacing. Next, grid points are evaluated, given their position on the protein and clustered in order to identify putative binding sites. The second approach covers the protein surface with dummy *spheres*, checks if they satisfy the given conditions and again, clusters the results. The final group of geometry-based algorithms utilizes *α-shape theory*. Here the protein structure is preprocessed using Delaunay triangulation/Voronoi diagrams and the pocket is identified based on a variety of filtering criteria.

In comparison to the geometry algorithms, energy-based algorithms instead of calculating favorable distances among sidechain atoms calculate the interaction energy between dummy spheres and sidechain atoms. These spheres are further clustered and ranked based on the energies. The top scoring clusters are in turn reported as favorable ligand binding pockets.

It is hard to define which of the highlighted approaches is the most suitable for binding site prediction, as they under or overestimate certain characteristics. Usually the best approach is to try a couple of them and select the most relevant result based on the consensus between different algorithms. This is the approach taken by the popular service MetaPocket [21], which is discussed in detail in Chap. 5 – Detection and Extraction of Fragments.

Below you can find an example of the successful application of this technique in the life-science domain.

2.2.1 Ubiquitin-Binding Domain Prediction

The family of small regulatory proteins – ubiquitin is responsible for a remarkable range of functions. Ubiquitin can be covalently attached to a specific substrate protein, the process is referred to as ubiquitination. Ubiquitination is responsible for the trafficking of endogenous and retroviral transmembrane proteins. Additionally, it was shown that the blocking of distinct ubiquitin binding domains (UBDs) in vivo can influence retroviral budding. Therefore, the successful identification of novel ubiquitin binding domains can contribute to the design of novel selective drugs. A database-wide study has been successfully conducted [22] in order to identify previously undiscovered UBDs. They found the apoptosis-linked gene 2 interacting protein X (ALIX) to contain a potential new UBD, specifically the central V domain. These in silico findings were later confirmed experimentally by biophysical affinity measurements.

2.2.2 Pattern Detection

In contrast to the prediction of protein structural patterns, there are software tools and approaches capable of their direct detection. The subtle difference between the two is rather simple. Prediction strives to make an educated guess as to whether or not an arrangement of amino acids will have the desired characteristics, while direct detection only identifies patterns with user-defined properties. For example you can specify a pattern composition at the atomic, residual or secondary structure level; restrict inter-atomic distances, or bond connections. This can be particularly useful for pharmacophore search and for the extraction of more general patterns of interest.

In the following section we review some of the tools used for pattern detection. RASMOT-3D PRO [23] is a web service performing systematic searches of 3D structures given a user-defined structural pattern. The pattern exploration is limited to up to 10 selected protein structures or a non-redundant set of PDB chains. An estimate of whether or not a found pattern corresponds to the query structure is made based on the comparison of C_α and C_β atoms altogether with the RMSD.[2] Another powerful service, which is directly incorporated into the Protein Data Bank in Europe [24] is PDBeMotif [25]. This web application allows a wide range of pre-defined search functions; however its customization is limited to the pre-defined parameters. Another drawback to this approach is the fact that the precomputed data in the database are stored for individual protein chains, therefore neglecting all patterns concerned with the interface of chains. In comparison, PatternQuery [26] is a language and a web-service covering the majority of the former search, taking into consideration the PDB entry as a whole. The advantage is that by using clear and highly customizable syntax, all the queries can be accurately tailored according to the user's needs, even covering complex patterns. More information on the functionality of PatternQuery is provided in Chap. 5. Finally, IMAAAGINE [27] is designed for the identification of patterns up to 8 amino acids (AA) in size with pre-defined distances, thus completely neglecting the bound ligands. Last but not least, ASSAM [28] identifies user-defined patterns of up to 12 AAs.

Below you can find an example of a pattern detection protocol successfully applied in the field of drug design.

2.2.3 Phosphorylation of Drug Binding Pockets

Roughly half of eukaryotic proteins are subject to a post-translational modification – phosphorylation. This addition of a phosphate group to certain amino acid residues can greatly influence the properties of a binding site which is subject to drug inhi-

[2]RMSD is a metric describing the structural difference between two molecules (patterns) in Ångströms, i.e. how well would two or more structures fit on top of each other. The higher the RMSD is, the more divergent the structures are. Two molecules with identical conformation (same atomic positions) have an RMSD equal to 0.

bition. A recent database-wide survey [29] examined mammalian proteins with the bound drug ligand. In particular, target-bound ligands together with residues within 12 Å of the binding site have been extracted and inspected for phosphorylation. Over 70 % (453) of the proteins exhibited phosphorylation. Almost one third of them (132) exhibited this phosphorylation in the vicinity of the binding site, and therefore can alter ligand binding. For 70 out of the 132 examples, it is known whether or not phosphorylation alters drug binding. 27 of them exhibited similar effects on activity even after phosphorylation, in contrast to the other 43, whose effects were the opposite.

For example, cyclin-dependent kinase 2 (CDK2) is an enzyme catalyzing the phosphoryl transfer of ATP phosphate group to serine or threonine hydroxyl in a protein substrate, a process important in cell cycle regulation. In particular, the enzyme exhibits phosphorylation both at a positive and negative regulatory site [30]. While the phosphorylation of threonine 160 in the vicinity of the active site activates the enzyme function [31], the phosphorylation of tyrosine 15 negatively affects substrate binding [32, 33].

This is just one example of how the database-wide identification, extraction and analysis of structural patterns can provide a fresh insight into the phosphorylation of an inhibitor's binding sites in the context of rational drug design. Using sophisticated tools like PatternQuery can tremendously simplify the complexities of obtaining input data for various types of analyses, and therefore enable analyses to be carried out that were not feasible before.

References

1. Daëron, M., Jaeger, S., Du Pasquier, L., Vivier, E.: Immunoreceptor tyrosine-based inhibition motifs: a quest in the past and future. Immunol. Rev. **224**(1), 11–43 (2008). doi:10.1111/j.1600-065X.2008.00666.x
2. Laskowski, R.A., Gerick, F., Thornton, J.M.: The structural basis of allosteric regulation in proteins. FEBS Lett. **583**(11), 1692–1698 (2009). doi:10.1016/j.febslet.2009.03.019
3. Motlagh, H.N., Wrabl, J.O., Li, J., Hilser, V.J.: The ensemble nature of allostery. Nature **508**(7496), 331–339 (2014). doi:10.1038/nature13001
4. Nussinov, R., Tsai, C.J.: Allostery in disease and in drug discovery. Cell **153**(2), 293–305 (2013). doi:10.1016/j.cell.2013.03.034
5. Liang, J., Woodward, C., Edelsbrunner, H.: Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. Protein Sci. **7**(9), 1884–1897 (1998). doi:10.1002/pro.5560070905
6. Nayal, M., Honig, B.: On the nature of cavities on protein surfaces: application to the identification of drug-binding sites. Proteins: Struct. Funct. Bioinf. **63**(4), 892–906 (2006). doi:10.1002/prot.20897
7. Skolnick, J., Gao, M., Roy, A., Srinivasan, B., Zhou, H.: Implications of the small number of distinct ligand binding pockets in proteins for drug discovery, evolution and biochemical function. Bioorg. Med. Chem. Lett. **25**(6), 1163–1170 (2015). doi:10.1016/j.bmcl.2015.01.059
8. Hubner, C.A.: Ion channel diseases. Hum. Mol. Genet. **11**(20), 2435–2445 (2002). doi:10.1093/hmg/11.20.2435
9. Zhou, H.X., McCammon, J.A.: The gates of ion channels and enzymes. Trends in Biochem. Sci. **35**(3), 179–185 (2010). doi:10.1016/j.tibs.2009.10.007

10. Smith, W.L., DeWitt, D.L., Garavito, R.M.: Cyclooxygenases: structural, cellular, and molecular biology. Ann. Rev. Biochem. **69**(1), 145–182 (2000). doi:10.1146/annurev.biochem.69.1.145
11. Hornak, V., Simmerling, C.: Targeting structural flexibility in HIV-1 protease inhibitor binding. Drug Discov. Today **12**(3–4), 132–138 (2007). doi:10.1016/j.drudis.2006.12.011
12. Kunze, J., Todoroff, N., Schneider, P., Rodrigues, T., Geppert, T., Reisen, F., Schreuder, H., Saas, J., Hessler, G., Baringhaus, K.H., Schneider, G.: Targeting dynamic pockets of HIV-1 protease by structure-based computational screening for allosteric inhibitors. J. Chem. Inf. Mod. **54**(3), 987–991 (2014). doi:10.1021/ci400712h
13. Pabo, C.O., Peisach, E., Grant, R.A.: Design and selection of Novel Cys 2 His 2 zinc finger proteins. Ann. Rev. Biochem. **70**(1), 313–340 (2001). doi:10.1146/annurev.biochem.70.1.313
14. Dundas, J., Ouyang, Z., Tseng, J., Binkowski, A., Turpaz, Y., Liang, J.: CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. Nucl. Acids Res. **34**(Web Server), W116–W118 (2006). doi:10.1093/nar/gkl282
15. Yu, J., Zhou, Y., Tanaka, I., Yao, M.: Roll: a new algorithm for the detection of protein pockets and cavities with a rolling probe sphere. Bioinformatics **26**(1), 46–52 (2010). doi:10.1093/bioinformatics/btp599
16. Laurie, A.T.R., Jackson, R.M.: Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. Bioinformatics **21**(9), 1908–1916 (2005). doi:10.1093/bioinformatics/bti315
17. Ngan, C.H., Hall, D.R., Zerbe, B., Grove, L.E., Kozakov, D., Vajda, S.: FTSite: high accuracy detection of ligand binding sites on unbound protein structures. Bioinformatics (Oxford, England) **28**(2), 286–7 (2012). doi:10.1093/bioinformatics/btr651
18. Sehnal, D., Svobodová Vařeková, R., Berka, K., Pravda, L., Navrátilová, V., Banáš, P., Ionescu, C.M., Otyepka, M., Koča, J.: MOLE 2.0: advanced approach for analysis of biomacromolecular channels. J. Cheminf. **5**(1), 39 (2013). doi:10.1186/1758-2946-5-39
19. Chovancova, E., Pavelka, A., Benes, P., Strnad, O., Brezovsky, J., Kozlikova, B., Gora, A., Sustr, V., Klvana, M., Medek, P., Biedermannova, L., Sochor, J., Damborsky, J.: CAVER 3.0: a tool for the analysis of transport pathways in dynamic protein structures. PLoS Comput. Biol. **8**(10), e1002,708 (2012). doi:10.1371/journal.pcbi.1002708
20. Yaffe, E., Fishelovitch, D., Wolfson, H.J., Halperin, D., Nussinov, R.: MolAxis: a server for identification of channels in macromolecules. Nucl. Acids Res. **36**(Web Server issue), W210–5 (2008). doi:10.1093/nar/gkn223
21. Huang, B.: MetaPocket: a meta approach to improve protein ligand binding site prediction. OMICS: J. Integr. Biol. **13**(4), 325–330 (2009). doi:10.1089/omi.2009.0045
22. Ehrt, C., Brinkjost, T., Koch, O.: Impact of binding site comparisons on medicinal chemistry and rational molecular design. J. Med. Chem. **59**(9), 4121–4151 (2016). doi:10.1021/acs.jmedchem.6b00078
23. Debret, G., Martel, A., Cuniasse, P.: RASMOT-3D PRO: a 3D motif search webserver. Nucl. Acids Res. **37**(SUPPL. 2), 459–464 (2009). doi:10.1093/nar/gkp304
24. Velankar, S., van Ginkel, G., Alhroub, Y., Battle, G.M., Berrisford, J.M., Conroy, M.J., Dana, J.M., Gore, S.P., Gutmanas, A., Haslam, P., Hendrickx, P.M.S., Lagerstedt, I., Mir, S., Fernandez Montecelo, M.A., Mukhopadhyay, A., Oldfield, T.J., Patwardhan, A., Sanz-García, E., Sen, S., Slowley, R.A., Wainwright, M.E., Deshpande, M.S., Iudin, A., Sahni, G., Salavert Torres, J., Hirshberg, M., Mak, L., Nadzirin, N., Armstrong, D.R., Clark, A.R., Smart, O.S., Korir, P.K., Kleywegt, G.J.: PDBe: improved accessibility of macromolecular structure data from PDB and EMDB. Nucl. Acids Res. **44**(D1), D385–D395 (2016). doi:10.1093/nar/gkv1047
25. Golovin, A., Henrick, K.: MSDmotif: exploring protein sites and motifs. BMC Bioinf. **9**, 312 (2008). doi:10.1186/1471-2105-9-312
26. Sehnal, D., Pravda, L., Svobodová Vařeková, R., Ionescu, C.M., Koča, J.: PatternQuery: web application for fast detection of biomacromolecular structural patterns in the entire protein data bank. Nucl. Acids Res. **43**(W1), W383–W388 (2015). doi:10.1093/nar/gkv561

27. Nadzirin, N., Willett, P., Artymiuk, P.J., Firdaus-Raih, M.: IMAAAGINE: a webserver for searching hypothetical 3D amino acid side chain arrangements in the protein data bank. Nucl. Acids Res. **41**(Web Server issue) (2013). doi:10.1093/nar/gkt431
28. Nadzirin, N., Gardiner, E.J., Willett, P., Artymiuk, P.J., Firdaus-Raih, M.: SPRITE and ASSAM: web servers for side chain 3D-motif searching in protein structures. Nucl. Acids Res. **40**(Web Server issue), W380–6 (2012). doi:10.1093/nar/gks401
29. Smith, K.P., Gifford, K.M., Waitzman, J.S., Rice, S.E.: Survey of phosphorylation near drug binding sites in the protein data bank (PDB) and their effects. Proteins: Struct. Funct. Bioinf. **83**(1), 25–36 (2014). doi:10.1002/prot.24605
30. Morgan, D.O.: CYCLIN-DEPENDENT KINASES: engines, clocks, and microprocessors. Ann. Rev. Cell Dev. Biol. **13**(1), 261–291 (1997). doi:10.1146/annurev.cellbio.13.1.261
31. Gu, Y., Rosenblatt, J., Morgan, D.O.: Cell cycle regulation of CDK2 activity by phosphorylation of Thr160 and Tyr15. EMBO J. **11**(11), 3995–4005 (1992). http://www.ncbi.nlm.nih.gov/pubmed/1396589. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC556910
32. Bartova, I.: The mechanism of inhibition of the cyclin-dependent kinase-2 as revealed by the molecular dynamics study on the complex CDK2 with the peptide substrate HHASPRK. Protein Sci. **14**(2), 445–451 (2005). doi:10.1110/ps.04959705
33. Otyepka, M., Bártová, I., Kříž, Z., Koča, J.: Different mechanisms of CDK5 and CDK2 activation as revealed by CDK5/p25 and CDK2/Cyclin a dynamics. J. Biol. Chem. **281**(11), 7271–7281 (2006). doi:10.1074/jbc.M509699200

http://www.springer.com/978-3-319-47387-1

Structural Bioinformatics Tools for Drug Design
Extraction of Biologically Relevant Information from Structural Databases

Koča, J.; Svobodová Vařeková, R.; Pravda, L.; Berka, K.; Geidl, S.; Sehnal, D.; Otyepka, M.

2016, XIII, 144 p. 62 illus., Softcover
ISBN: 978-3-319-47387-1