Performance of basic text mining for specific protein-protein complexes

SH2D1A – p59Fyn complex (1m27). For this complex, AND-query did not retrieve any abstracts. The OR-query identified 6 abstracts with 4 residues, out of which 3 are at the interface (Figure 3 in the main text, \( P_{TM} = 0.75 \)). Arg78 of SH2D1A protein (1m27, chain A) was detected in the abstract on the role of tyrosine kinase Fyn and SLAM (synonym of SH2D1A) interaction in the development of natural killer T cells in human and mice [1]. The abstract was not detected by AND-query because p59Fyn names that do not follow UniProt nomenclature. Arg78 was also pinpointed in two mutagenesis studies on the role of the SLAM-SAP-Fyn signaling pathway in mice CD4 T cell function and germinal center development [2, 3]. Trp119 of p59Fyn (1m27, chain C) was detected in the abstract on interactions of SH2 and SH3 domains of p59Fyn [4].

TACE – TIMP3 complex (3cki). For human tumor necrosis factor alpha converting enzyme (TACE) co-crystallized with the metalloproteinase inhibitor 3, TIMP3, the AND-query predicted 7 interface and 5 non-interface residues (Figure S1A, \( P_{TM} = 0.58 \)), mentioned in the abstracts of 6 publications [7-12], other than the original X-ray study [13]. Redesign of metalloproteinase inhibitor TIMP1, using TIMP3 as a scaffold identified three residues, Ser4, Leu67, and Arg84, important for the TIMP3-TACE binding [7]. Ser4 and Thr2 were found to be functionally important by measuring binding affinities of the mutated TIMP3 [8]. Redesign of metalloproteinase inhibitor TIMP2, using TIMP3 and TIMP1 as scaffolds [9], identified Ser4 and, additionally, Phe34 residues. Three other residues mentioned in that abstract (Val/Leu69, Thr/Leu98 and Arg100, respectively) did not match residues in the original X-ray crystallography paper [13]. In the study of TACE cysteine-rich domains role in the TIMP-3 inhibitory potency [10], Lys315 was mentioned as a residue adjacent to the TACE catalytic site, which interacts with TIMP-3 Glu31, which is close to, but not at the TACE-TIMP3 interface. In this abstract, other non-interface residues, Lys26, Lys27, Lys30, and Lys76 were also mentioned along with residues (Glu26, Glu27, Glu30, and Glu76) that either did not match PDB numbering or did not follow residue patterns considered in this study (Table 1). Finally, Pro5 is identified in the analysis of expression levels of different human prostatic tumor cell lines [11]. However, Pro5 in this publication stands for the name of cell line rather than a residue. Ser4 was also identified by another publication [9] where the binding affinity of the mutant is better than TIMP-3. All the above residues belong to TIMP3 chain in the original PDB, except Lys315, which belongs to the TACE chain.

The OR-query for this complex found three additional abstracts [14-16] with two additional interface residues (TACE Val353 and TIMP3 Ala11) and one non-interface (TIMP3 Gly119) residue \( (P_{TM} = 0.60) \). Val353 was identified as functionally important in the study on stabilization of the TACE autoproteolysis [14]. Ala11 is mentioned in the abstract of paper [15] showing that activation of brain ET(B) receptors causes TIMP-1 and TIMP-3 production. Gly119 is present in the abstract of study [16] on mutated
growth hormone (bGH) in transgenic mice. This abstract was picked due to the presence of the TIMP-3 name, however Gly119 mentioned therein belongs to bGH protein and just accidentally coincides with the Gly119 residue in the original PDB file. The OR-query also found three additional abstracts with five residues already picked up by the AND-query. Study [17] determined that Lys26, Lys27, Lys30, and Lys76 constitute another TIMP3 binding site with the extracellular matrix (ECM). Other residues mentioned in that abstract (Arg163 and Lys165) belong to the C-terminal of TIMP3 not present in the original PDB file. Lys27 was also present in the abstract of study [18] suggesting that EZH2 (Enhancer of zeste homolog 2) accelerates lung cancer cell migration partially through repression of TIMP-3 expression. Thr2 was identified in the paper [19] showing that mutated N-TIMP3 inhibits degradation of ADAMTS-4 and ADAMTS-5 metalloproteinases.

**Complex of Plectin-1 and Integrin beta-4 (3f7p).** For this complex, AND-query found 3 abstracts with 4 interface residues of Integrin beta 4 (Figure S1B, $P_{TM} = 1.00$). Abstract in Ref. [20] states that mutations at Arg1225 and Arg1281 sites inhibit interaction of the integrin beta 4 with plectin while mutation of Lys1279 has no effect to recruit plectin. In addition, Arg1281 was spotted in the abstract in Ref. [21] also showing that mutation at this site affects the interaction between the two proteins. Another mutagenesis study [22] shows that residue Ser1325 is important for recruitment of plectin into hemidesmosomes in vivo.

The OR-query for this complex identified one additional abstract with one additional non-interface residue (Figure S1B) thus reducing TM performance to $P_{TM} = 0.80$. Arg239 was found in the abstract in Ref. [23] on interaction between plectin and glial fibrillary acidic protein (GFAP). However, the number of this residue belonging to GFAP protein, just accidentally coincides with the number of one of the arginine residues in the original PDB file. The OR-query has also retrieved 2 additional abstracts of mutagenesis studies containing Arg1281 pinpointed by the AND-query. Mutation at Arg1281 was found to affect severity of epidermolysis bullosa (genetic skin disease) [24] and to inhibit interaction between plectin and integrin beta 4 with its alpha 6 chain [25].

**IRF3 – CBP complex (1zoq).** For this complex, the AND-query identified 6 abstracts with 3 non-interface residues in the Interferon regulatory factor 3, IRF3 (Figure S1C, $P_{TM} = 0.00$). All found residues (Ser339 [26, 27], Ser385 [28-30], and Ser386 [28-31]) were studied in the context of their phosphorylation, which regulates CBP binding allosterically. Interferon interacts with many partners (e.g., BioGrid database [32] lists 44 interactions for IRF-3) and phosphorylation of these three residues was studied in the context of IRF3 binding to other proteins as well. The OR-query thus found Ser339, Ser385 and Ser386 residues in one [33], two [34, 35], and five [35-39] additional abstracts, respectively.

The OR-query for this complex found one more residue, Leu238 of IRF3, in one additional abstract ($P_{TM} = 0.00$) of study on the interferon (IFN) role in resisting evasion of the African swine fever virus into pig immune system [40].
References


