J Mol Model (2007) 13:445–455
DOI 10.1007/s00894-006-0165-4

ORIGINAL PAPER

LigProf: A simple tool for in silico prediction of ligand-binding sites

Grzegorz Koczyk · Lucjan S. Wyrwicz · Leszek Rychlewski

Received: 27 June 2006 / Accepted: 25 October 2006 / Published online: 3 January 2007

Abstract With the increasing amount of data provided by both high-throughput sequencing and structural genomics studies, there is a growing need for tools to augment functional predictions for protein sequences. Broad descriptions of function can be provided by establishing the presence of protein domains associated with a particular function. To extend the domain-based annotation, LigProf provides predictions of potential ligands that bind to a protein, as well as critical residues that stabilize ligands. A P-value statistic for estimating the significance of motif occurrence is provided for all sites. Although the usefulness of the method will rise with increasing numbers of crystallographically solved molecules deposited in the PDB database, we show that it can already be applied successfully to the highly represented ligand-bound protein kinase domains of viral and human origin. The LigProf webserver is freely available at: http://www.cropnet.pl/ligprof. At present, LigProf descriptors annotate and extend major protein families from the PfamA database.

G. Koczyk (✉)
Institute of Plant Genetics,
Strzeszyńska 34,
60-479 Poznań, Poland
e-mail: gkoc@igr.poznan.pl

G. Koczyk · L. S. Wyrwicz · L. Rychlewski
BioInfoBank Institute,
Limanowskiego 24A,
60-744 Poznań, Poland

L. S. Wyrwicz
Department of Gastroenterology, Medical Center
for Postgraduate Education and Maria Sklodowska-Curie
Memorial Cancer Center and Institute of Oncology,
Roentgena 5,
02-781 Warsaw, Poland

Keywords Ligand-binding sites · Profiles · Structural genomics

Introduction

The automated assignment of function to protein sequences has long been a subject of interest in bioinformatics, [1] in direct proportion to the abundance of data provided by new genomic projects. Quickly growing numbers of sequences in databases such as the NCBI GenBank make it extremely unlikely for even a cursory in vitro study of most proteins' activities. As a consequence, annotations are usually derived from homology to sequences of known function. In particular, in silico prediction of biocatalytic activity is desirable from both the theoretical (as a classification problem) and the applied (biocatalysts for pharmaceutical and chemical industry sectors - [2]).

There are two important questions that sequence analysis can try to answer, in order to satisfy the needs of researchers (modified after [2]):

(1) How to efficiently predict proteins from various functional classes ? ("How can functional homology best be determined to identify suitable donor enzymes?")
(2) What residues are functionally important ? ("What is the best way to choose amino acid substitutions from these donors ?" [e.g., for directed mutagenesis])

For answering the first question, 'best hit' predictions that use a sequence-database-searching tool such as BLAST or PSI-BLAST [3] are frequently overused. Analyses published by Devos and Valencia [4] and Rost and co-workers [5] estimate that from 10 to 30% of functional annotations may be erroneous, using Enzyme Classification numbers as a

criterion. This is mostly due to overly specific descriptions in cases where no strong evidence exists. However, it is worth noting that an intrinsic, unavoidable annotation error can arise from the ambiguities in the Enzyme Classification itself [6].

Establishing the presence of significant homology to a particular protein fold/family represents a safer alternative for broad functional prediction (with reliable transfer up to the third level of Enzyme Classification [7]). In many cases, this can be accomplished quickly by scanning against a protein family database, such as: Pfam [8], SMART [9] or NCBI's Conserved Domains Database [10]. However, the exact substrate and cofactor specificities (fourth level of Enzyme Classification) are not determined reliably by inference on the basis of overall homology [7].

At this stage, the prediction of function ties neatly into the second question (establishment of functionally important residues). The protein molecule's specificity is strongly influenced by only a fraction of protein's residues, a point that has been demonstrated for distant homologs with preserved molecular function [11]. The fraction consists largely of those parts of the sequence that form the active site (s) and/or cofactor-binding sites (which we will subsequently denote as either: 'ligand-binding site' or 'distributed motif').

Elucidation of functional sites from known protein structures and a subsequent search for similar spatial patterns have already provided a number of confident, functional predictions for several crystallized proteins of unknown function (e.g., the recently developed SiteBase -, [12] MultiBind - [13]). Analysis of the predicted sites' mode of action has already enabled successful rational drug design processes (e.g., influenza neuraminidase inhibitors, [14] adenosine deaminase inhibitors [15]).

Possible approaches to functional site description for new protein sequences center on modeling protein structure, followed by ligand docking (e.g., [16, 17]). The modeling approach is well suited to detection of novel specificities for existing protein targets, but its computational cost may be prohibitive for functional annotation in genomics projects. However, almost no up-to-date services (with a notable, recent exception of SMID-BLAST - [18]) are available for the direct annotation of protein sequences. Consequently, we suggest that there is a need for reliable structure-orientated ligand-binding-site-prediction services on the basis of sequence information with embedded statistical significance assessment. We propose a simple tool to fill that need in the form of our LigProf prediction service.

Methods

The residues that form a ligand-binding site are frequently distributed in sequence, but form a distinct spatial neighborhood of the ligand (see Fig. 1). A total of 35,087 protein structures from the PDB database [19] was downloaded and subsequently annotated with the LPC software [20] to establish residues in contact with small molecule ligands, and the most probable nature of interactions (categorized into: hydrophilic, hydrophobic, aromatic ring stacking and destabilizing interactions). Side-chain interactions are taken into account as possible specificity determinants. For destabilizing interactions, only those residues judged by LPC to have at least one of the other three types of interaction were retained. However, it is possible that neighboring residues can act as steric hindrances, forcing volume constraints on interacting residues despite a lack of side-chain interactions. The disruption of secondary structure is particularly strong in the case of proline residues, which we also include in LigProf descriptor construction, regardless of the nature of the interaction.

The residues were subsequently mapped onto the consensus sequence positions of PfamA families by matching with the *hmmsearch* program [21] to provide a reference frame for other steps of the method. It is worth noting that the mapping can result in a partial loss of information if some binding-site residues constitute part of a highly variable loop that is not incorporated in the consensus. As it happens, this is the case more often in protein-protein interactions (e.g., immunoglobulins) than in protein-small molecule interactions [22].

The motifs for each ligand are then clustered employing a complete-linkage hierarchical clustering algorithm, [23] with 50% overlap as a guiding criterion. Clustered motif instances are subsequently merged to construct a non-redundant set of motifs.

Lastly, each non-redundant motif is converted into a position-specific scoring matrix (i.e., profile [24]), with amino-acid occurrence in subsequent motif positions scored independently according to the best score obtained from an isomorphic substitution matrix [25]. The LigProf best score is defined as a maximum score of an amino acid (in query sequence) to any of the residues occurring at the position in those reference PDB structures that contributed to the motif.

To provide a P-value distribution for non-redundant motifs, the following generalizing assumptions must be made.

First, we assume that a particular set of protein residues capable of forming a binding site is under positive selection if and only if it occurs in a protein that has a functional site used to bind and/or process the ligand. Therefore a site indeed constitutes a motif that is conserved due to evolutionary constraints imposed by protein function (e.g., binding or catalysis). The same underlying assumption is commonly made in the detection of functionally important residues from sequence conservation (hypervariation of residues responsible for specificity and conservation of residues responsible for the overall conformation of the binding site - [26]) or via the "evolutionary trace" method (residues responsible for spec-

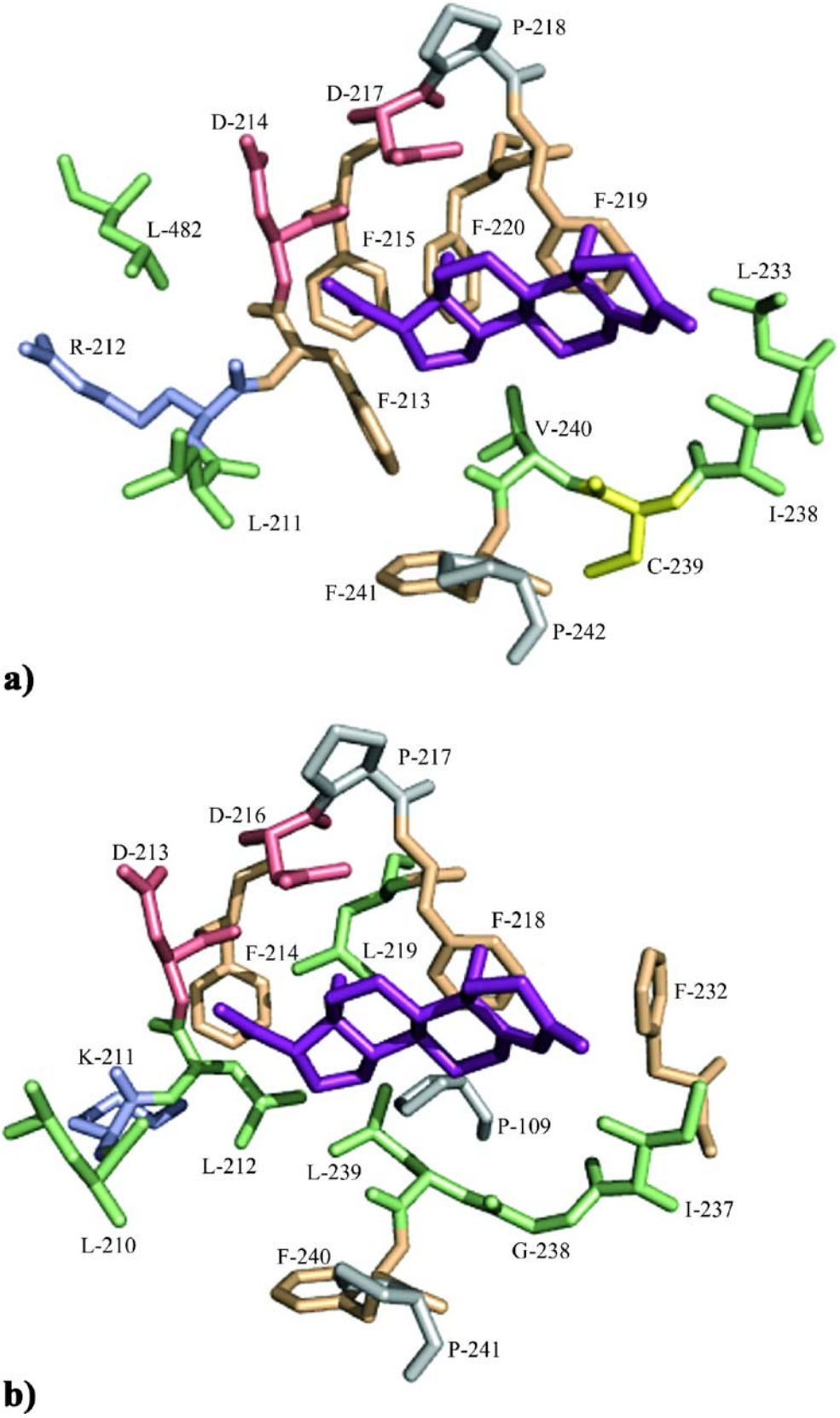


Fig. 1 Visualization of 7Å radius neighborhood for the ligand progesterone (magenta) **(a)** taken from the 1W0F reference protein structure, which provided the binding site descriptor **(b)** modeled for the query sequence using MODELLER (query the same as analyzed in Fig. 2)

ificity are conserved in the subtree corresponding to the subfamily of given specificity - [27, 28]). The higher conservation of binding residues was originally observed by Johnson and Church [29] and subsequently corroborated as a general trend for small-molecule ligands by Pils and co-workers [22].

The second assumption is that when considering an entire set of distant homologs from the family for the presence of a site, a majority of proteins do not interact with the molecule (have either different specificity or different mode of binding). Without functional constraints acting to preserve the binding site, residues capable of forming the distributed motif are under neutral selection.

In order to provide P-value estimates of residues forming a high-scoring false binding site, we estimate the neutrally selected 'background' distributions of amino-acid frequencies for each position of the ligand-binding site. Datasets for estimation consist of family homologues detected by 3 iterations of PSI-BLAST [3] (with an E-value threshold for the inclusion set to 0.001) on the UniProt protein database, [30] clustered at 70% sequence identity with the CD-HIT program [28].

A simple, but effective (as shown in PSI-BLAST itself [3]) position-independent weighting scheme [31] is used during frequency estimation to protect against over-emphasizing groups of highly similar sequences. Frequency estimates are then provided by a substitution-matrix based regularizer (outlined in [3]).

Using the scoring function (values from the profile) and the estimated background distribution of amino acids, the entire probability distribution function and the corresponding P-values can be enumerated exhaustively according to the following set of equations: [32]

$$f^{(0)}(x) = \delta(x)$$

$$\forall_{j\in(1,2,3,\ldots,J)} f^{(j)}(x) = \sum_{a\in A} q_j(a) f^{(j-1)}\left(x - S_j(a)\right)$$

$$f(x) = f^{(J)}(x)$$

$$P(x) = \sum_{x'\geq x} f(x')$$

where:

- A motif alphabet (in this case: 20 amino-acid residues, plus symbols for undefined residue and gap),
- $\delta(x)$ initial value of the probability (1.0 for x=0 and 0.0 otherwise),
- $f(x)$ total probability of obtaining score x for a match across all the binding site positions,
- $f^{(j)}(x)$ probability of obtaining score x for a match up to and including motif position j,
- J total number of positions that constitute the 'merged' binding site,
- $P(x)$ total probability of obtaining a score no worse than x for a match across all the binding-site positions (match P-value),
- $q_j(a)$ frequency of letter a in position j,
- $S_j(a)$ score of letter a in position j of the motif.

An example output for a LigProf profile match is summarized in Fig. 2, the motif creation pipeline is shown in Fig. 3.

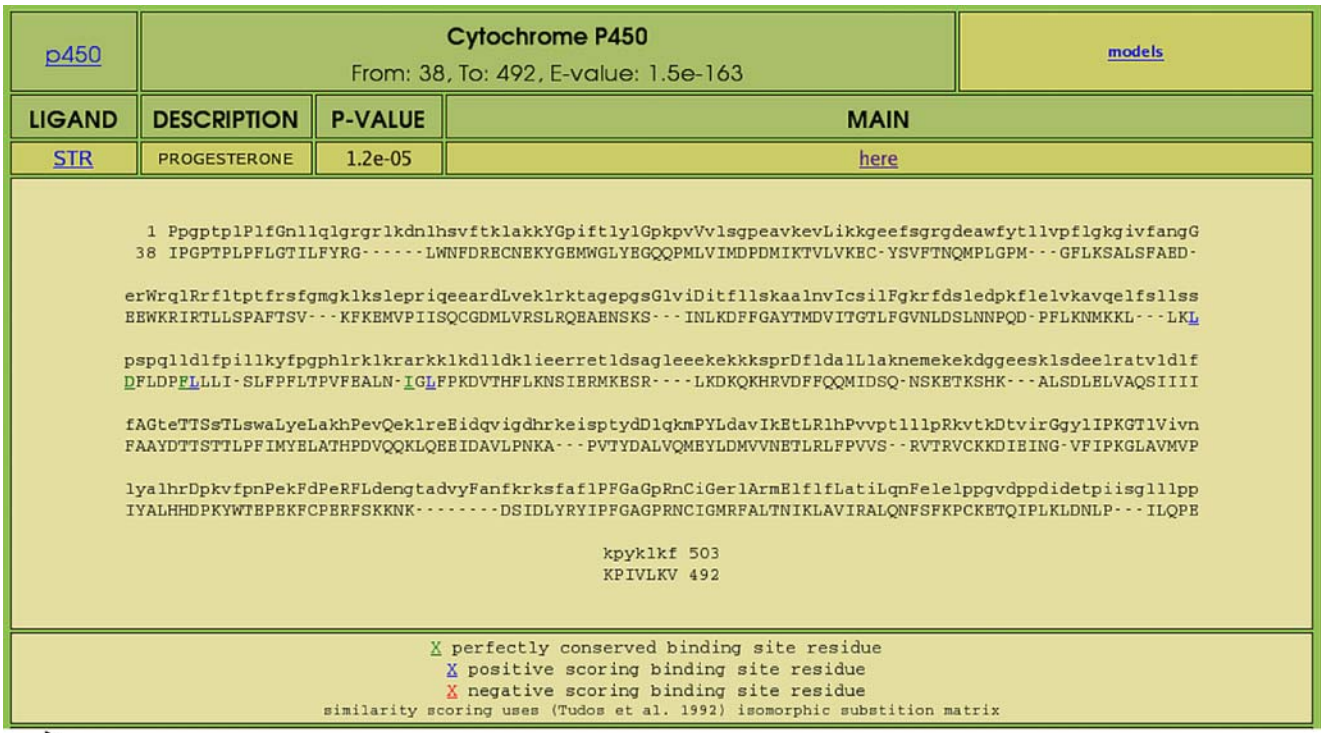
To enhance the usability of LigProf further, we provide an option to create homology-based models using MODELLER [33] on the basis of the domain alignment provided by the *hmmalign* program [21]. While not the best of choices, *hmmalign* performs reasonably well in terms of multiple sequence-alignment quality [34] and computation resources required. LigProf provides the following: a model of the entire domain instance (in the query) and models of the site in both the query and template structures. Example site models for query and template P450 cytochrome sequences are shown in Fig. 1a and b respectively. These correspond to a binding site shown in the LigProf example output (Fig. 2).

The binding site models are intended to provide a quick way to inspect the entire spatial neighborhood within a radius of 7.0 Å of the molecule. This is pertinent as some functionally important (e.g., catalytic) residues can have a destabilizing influence on the substrate (these amino acids are discarded during profile generation). The upper limit of 7 Å for specificity-determining residues has been noted before, for example in analyses of the LacI bacterial transporter family [35]. Due to the requirements imposed by MODELLER's academic license, the modeling feature is only available to registered MODELLER users upon entering the program's registration key.

The entire LigProf pipeline has been implemented in the Python/C++ programming languages. For parsing PDB files and construction of ligand-neighborhood models, LigProf makes use of the Bio.PDB library [36].

Results

One of the principal test cases we used to establish the usefulness of LigProf was the tryptophanylo/tyrosynylo-tRNA synthetase family. The evolutionary scenarios reconstructed for the family in [37] advocate a late development of discriminatory mechanisms for these two substrates (there is a relatively low sequence divergence between the two specificity's representative sets as compared to the divergence within the sets). Therefore, the family represents a well-suited test set for in silico predictions of specificity. Indeed, a set of family homologues has been used as a benchmark for the recent prediction service: SMID-BLAST ([18]; http://smid.blueprint.org).



a)

Motif information (contacts)

QUERY: POSITION	RESIDUE	HMM: POSITION	SOURCES	RESIDUES	Minimum DISTANCE (A)	Maximum contact SURFACE (A^2)
212	L	196	1W0F_A	F	3.2	29.4
213	D	197	1W0F_A	D	3.0	41.7
218	F	202	1W0F_A	F	3.7	54.7
219	L	203	1W0F_A	F	3.5	16.6
237	I	223	1W0F_A	I	4.1	27.1
239	L	225	1W0F_A	V	3.6	33.9

b)

Fig. 2 Example LigProf output corresponding to the ligand binding site of the ligand progesterone from Fig. 1a **(a)** alignment of query sequence (cytochrome 3A4) to domain consensus sequence, with underlined residues of the binding site colored according to score **(b)** reference information about the binding site, as represented in template structure

We used LigProf to predict the most significant binding sites for sets of 141 Tyr-tRNA synthetases and 172 Trp-tRNA synthetases. The sequences were obtained from scanning the NCBI non-redundant (nr) protein sequences database, clustered at 90% sequence identity with CD-HIT, [38] with the PfamA *tRNA_synt_1b* (PF00579) Hidden Markov Model. Results of best ligand prediction are summarized in Tables 1 and 2 for Tyr-tRNAand Trp-tRNA synthetases, respectively.

With the exception of three cases in the Tyr-tRNA synthetase non-redundant set, all sequences were predicted to have strongest binding sites for either potent artificial inhibitors (545 and 485 compounds [39]) or modified substrate (TYA, YSA). Our results for Tyr-tRNA synthetases are slightly better than those quoted for SMID-BLAST in [18], where 13 out of 83 (15%) sequences had tryptophanylo-5′-AMP (TYM) predicted as the best substrate.

The most promising area for LigProf predictions is initial evaluation of lead compounds by structural determinants of the active site. Therefore, to assess the utility of tool for aiding drug design, we queried LigProf with the sequences of divergent protein kinases of unknown specificity, searching for potential inhibitors. For the tests we selected viral proteins from the human pathogen *Herpesviridae* -

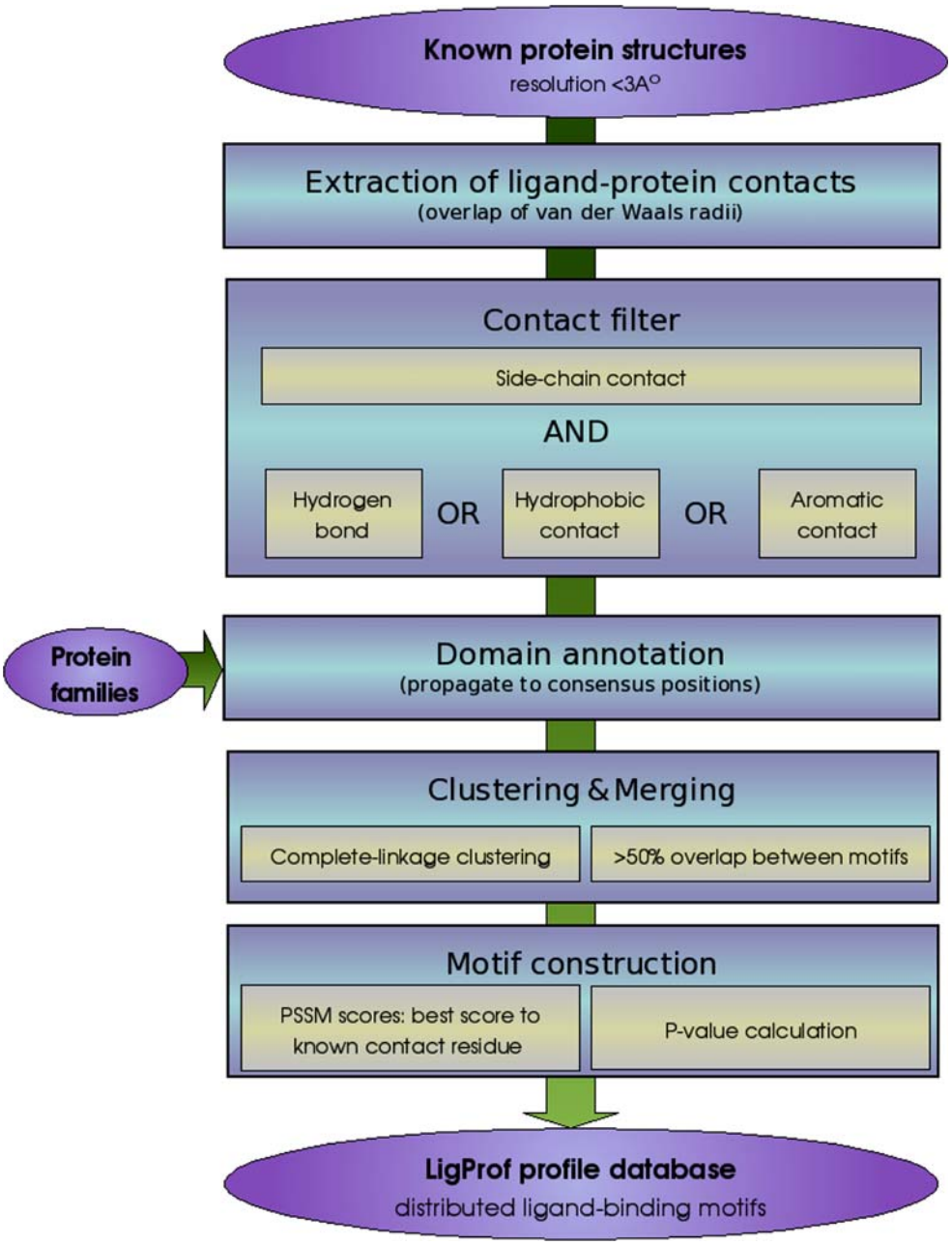


Fig. 3 Information flow in LigProf database construction pipeline

HSV1 serine/threonine kinases UL13, US3 and CMV UL97. These kinases represent potential targets for antiviral drug development as US3 kinase is critical for the inhibition of apoptosis in infected cells, [40] UL13 is known to regulate US3 activity [41] and CMV UL97 plays a critical role in viral replication and virion morphogenesis [42]. Sequences of the selected proteins were used to query the dataset of 215 inhibitors (243 distributed motifs), co-crystallized in 248 PDB kinase structures. The hits with P-values less than 0.05 are summarized in Tables 3, 4 and 5. It is worth noting that for each protein tested we observed a common pattern of inhibitory activity. Four P38 kinase ligands were identified for US3, while CDK2 ligands were identified for both UL13 and CMV UL97 kinase. In addition, the identified potential ligands of US3 contained several common pharmacophores (Fig. 4).

Discussion

We have designed and implemented a novel service for predicting binding sites on the basis of sequence similarity in residues determined from known structures. Treatment of a binding site as a distributed, conserved motif has enabled us to estimate the significance of the results in a clear way without the need of artificial, heuristic scoring [18]. Using the P-value estimates for binding sites, LigProf has been successful in assigning substrate specificities to members of

Table 1 Best binding sites predicted for Tyr-tRNA synthetases testing set

Abbreviation	Full name	Tyr-tRNA synthetase hits
YSA	5′-O-[N-(L-TYROSYL) SULFAMOYL]ADENOSINE	105
TYA	PHOSPHORIC ACID 2-AMINO-3-(4-HYDROXY-PHENYL)-PROPYL ESTER ADENOSIN-5′ YL ESTER	14
545	[2-AMINO-3-(4-HYDROXY-PHENYL)-PROPIONYLAMINO]-(1,3,4,5-TETRAHYDROXY-4-HYDROXYMETHYL-PIPERIDIN-2-YL)-ACETIC ACID BUTYL ESTER	12
TYB	TYROSINAL	3
TYM	**TRYPTOPHANYL-5′AMP**	**2**
485	[2-AMINO-3-(4-HYDROXY-PHENYL)-PROPIONYLAMINO]-(3,4,5-TRIHYDROXY-6-METHYL-TETRAHYDRO-PYRAN-2-YL)- ACETIC ACID	2
TYR	L-TYROSINE	2
TRP	**L-TRYPTOPHAN**	**1**

False negative hits shown in bold font

a non-redundant, representative subset of tryptophanylo/tyrosynylo-tRNA synthetases, as described in the Results section.

Because the coverage of the majority of enzyme families in PDB is low, [43] accurately modeling key residues of the active site that are responsible for the specificity of many subfamilies is not yet possible. To demonstrate the usefulness of the current version of LigProf for sequences with well-characterized domains, we screened the divergent viral kinases from *Herpesviridae* for potential inhibitors and observed a consistent specificity of ligands identified for each kinase. Based on the results of the analysis, we concluded that certain CDK2 inhibitors may interact with HSV1 UL13 kinase and CMV protein kinase UL97. This observation has found strong support in functional studies

Table 2 Best binding sites predicted for Trp-tRNA synthetases testing set

Abbreviation	Full name	Trp-tRNA synthetase hits
TYM	TRYPTOPHANYL-5′AMP	113
TRP	L-TRYPTOPHAN	31
LTR	L-TRYPTOPHAN	10
LTN	L-TRYPTOPHANAMIDE	8

Table 3 The summary of identified potential ligands of HSV US3 kinase, excluding substrates (ATP and its analogs), ions and nonspecific binders (e.g., glycerate)

ID	Name	P-value	Kinase
BMU	1-(5-TERT-BUTYL-2-METHYL-2H-PYRAZOL-3-YL)-3-(4-CHLORO-PHENYL)-UREA	0.008	P38
LI3	3 - FLUORO - N - 1H - INDOL - 5 - YL - 5 - MORPHOLIN - 4 - YLBENZAMIDE	0.013	P38
L09	N - (3 - TERT - BUTYL - 1H - PYRAZOL - 5 - YL) - N′ - {4 - CHLORO - 3 - [(PYRIDIN - 3 - YLOXY)METHYL]PHENYL}UREA	0.028	P38
L10	N-[(3Z)-5-TERT-BUTYL-2-PHENYL-1,2-DIHYDRO-3H-PYRAZOL-3-YLIDENE]-N′-(4-CHLOROPHENYL) UREA	0.030	P38

Human kinase type associated with the inhibitor is shown in last column.

of the influence of CDK2 inhibitors on gene expression of viral genes, [44] while the exact molecular mechanisms of the action still remain unknown for both UL13 [45] and UL97 [42]. The regulatory US3 kinase also exhibited a consistent pattern of potential inhibitors, with LigProf predicting affinity toward four ligands that have been co-crystallized previously with P38 kinase (Fig. 5) and share a common pattern of pharmacophores (Fig. 4). Notably, the tested proteins did not exhibit close homology to either CDK2 or P38 kinases, making the predicted specificities difficult targets for traditional annotation based on overall domain sequence similarity.

As the above example demonstrates, an effective application of LigProf is already possible. In silico annotation of divergent proteins (which cannot be mapped easily onto known lead compounds) enables us to draw conclusions as to the substrate/inhibitor specificities, provided the sequence belongs to a highly represented protein family (in terms of available structures of proteins co-crystallized with ligands).

Table 4 The summary of identified potential ligands of CMV UL97 kinase, excluding substrate (ATP and its analogs), ions and nonspecific binders (e.g., glycerate)

ID	Name	P-value	Kinase
CK1	4-(2,5-DICHLOROTHIEN-3-YL) PYRIMIDIN-2-AMINE	0.020	CDK2
N5B	N-(5-CYCLOPROPYL-1H-PYRAZOL-3-YL)BENZAMIDE	0.045	CDK2

Human kinase type associated with the inhibitor is shown in last column.

Table 5 The summary of identified potential ligands of HSV UL13 kinase, excluding substrate (ATP and its analogs), ions and non-specific ligands (e.g., glycerate)

ID	Name	P-value	Kinase
FBL	(2S)-1-[4-({4-[(2,5-DICHLOROPHENYL)AMINO]PYRIMIDIN-2-YL}AMINO)PHENOXY]-3-(DIMETHYLAMINO)PROPAN-2-OL	1.7e-04	CDK2
HDT	4-[(4-IMIDAZO[1,2-A]PYRIDIN-3-YLPYRIMIDIN-2-YL)AMINO]BENZENESULFONAMIDE	0.001	CDK2
106	4-(5-BROMO-2-OXO-2H-INDOL-3-YLAZO)-BENZENESULFONAMIDE	0.001	CDK2
CT9	4-[5-(TRANS-4-AMINOCYCLOHEXYLAMINO)-3-ISOPROPYLPYRAZOLO[1,5-A]PYRIMIDIN-7-YLAMINO]-N,N-DIMETHYLBENZENESULFONAMIDE	0.002	CDK2
U32	4 - [(5 - ISOPROPYL - 1,3 - THIAZOL - 2 - YL)AMINO]BENZENESULFONAMIDE	0.002	CDK2
4SP	O6-CYCLOHEXYLMETHOXY-2-(4′-SULPHAMOYLANILINO) PURINE	0.002	CDK2
I1P	2-[4-(N-(3-DIMETHYLAMINOPROPYL)SULPHAMOYL) ANILINO]-4-(IMIDAZO[1,2-B] PYRIDAZIN-3-YL)PYRIMIDINE	0.003	CDK2
CK6	4-[4-(4-METHYL-2-METHYLAMINO-THIAZOL-5-YL)-PYRIMIDIN-2-YLAMINO]-PHENOL	0.003	CDK2
LS3	3-{[(2,2-DIOXIDO-1,3-DIHYDRO-2-BENZOTHIEN-5-YL)AMINO]METHYLENE}-5-(1,3-OXAZOL-5-YL)-1,3-DIHYDRO-2H-INDOL-2-ONE	0.004	CDK2
CK9	2 - {[(2 - {[(1R) - 1 - (HYDROXYMETHYL)PROPYL]AMINO} - 9 - ISOPROPYL - 9H - PURIN - 6 - YL)AMINO]METHYL}PHENOL	0.004	CDK2
LS4	4-{[(2-OXO-1,2-DIHYDRO-3H-INDOL-3-YLIDENE)METHYL]AMINO}-N-(1,3-THIAZOL-2-YL) BENZENESULFONAMIDE	0.004	CDK2
628	4-{[6-(2,6-DICHLOROBENZOYL)IMIDAZO[1,2-A]PYRIDIN-2- YL]AMINO} BENZENESULFONAMIDE [PHENYLAMINOIMIDAZO(1,2-ALPHA)PYRIDINE]	0.005	CDK2
INR	2′,3-DIOXO-1,1′,2′,3-TETRAHYDRO-2,3′-BIINDOLE-5′-SULFONIC ACID	0.005	CDK2
LS1	N-METHYL-4-{[(2-OXO-1,2-DIHYDRO-3H-INDOL-3-YLIDENE)METHYL]AMINO} BENZENESULFONAMIDE	0.006	CDK2
RRC	R-ROSCOVITINE	0.006	CDK5
RYU	(2E,3S)-3-HYDROXY-5′-[(4-HYDROXYPIPERIDIN-1-YL)SULFONYL]-3-METHYL-1,3-DIHYDRO-2,3′-BIINDOL-2′(1′H)-ONE	0.009	CDK2
514	(3Z)-5-ACETYL-3-(BENZOYLIMINO)-3,6- DIHYDROPYRROLO[3,4-C]PYRAZOL-5-IUM	0.012	CDK2
STU	STAUROSPORINE	0.013	CSK, MAP2K2, CHK1, LCK
D31	2 - (4 - (AMINOMETHYL)PIPERIDIN - 1 - YL) - N - (3_CYCLOHEXYL - 4 - OXO - 2,4 - DIHYDROINDENO[1,2 - C]PYRAZOL - 5 - YL)ACETAMIDE	0.014	CDK2
ALH	6-PHENYL[5H]PYRROLO[2,3-B]PYRAZINE	0.015	CDK5
ST8	4-{[4-AMINO-6-(CYCLOHEXYLMETHOXY)-5-NITROSOPYRIMIDIN-2-YL]AMINO}BENZAMIDE	0.017	CDK2
HMD	4-(5-AMINO-4-OXO-4H-PYRAZOL-3-YL)-2-BROMO-4,5,6,7-TETRAHYDRO-3AH-PYRROLO[2,3-C]AZEPIN-8-ONE	0.020	CDK2
LS2	N-METHYL-{4-[2-(7-OXO-6,7-DIHYDRO-8H-[1,3]THIAZOLO[5,4-E]INDOL-8-YLIDENE) HYDRAZINO]PHENYL}METHANESULFONAMIDE	0.023	
CT7	(5-CHLOROPYRAZOLO[1,5-A]PYRIMIDIN-7-YL)-(4-METHANESULFONYLPHENYL)AMINE	0.023	CDK2
DT4	4 - ({5 - [(4 - AMINOCYCLOHEXYL)AMINO][1,2,4]TRIAZOLO[1,5 - A]PYRIMIDIN - 7 - YL} AMINO)BENZENESULFONAMIDE	0.024	CDK2
IXM	(Z)-1H,1′H-[2,3′]BIINDOLYLIDENE-3,2′-DIONE-3-OXIME	0.025	CDK5
N5B	N-(5-CYCLOPROPYL-1H-PYRAZOL-3-YL)BENZAMIDE	0.027	CDK2
1PU	1-(5-OXO-2,3,5,9B-TETRAHYDRO-1H-PYRROLO[2,1-A]ISOINDOL-9-YL)-3-PYRIDIN-2-YL-UREA	0.028	CDK2, CDK4
DTQ	4-[3-HYDROXYANILINO]-6,7-DIMETHOXYQUINAZOLINE	0.028	CDK2
D23	6-(3-AMINOPHENYL)-N-(TERT-BUTYL)-2-(TRIFLUOROMETHYL) QUINAZOLIN-4-AMINE	0.029	CDK2
BWP	(2S)-1-{4-[(4-ANILINO-5-BROMOPYRIMIDIN-2-YL)AMINO]PHENOXY}-3-(DIMETHYLAMINO) PROPAN-2-OL	0.033	CDK2
HDU	N-[4-(2-METHYLIMIDAZO[1,2-A]PYRIDIN-3-YL)-2-PYRIMIDINYL]ACETAMIDE	0.035	CDK2
FSE	3,7,3′,4′-TETRAHYDROXYFLAVONE	0.046	

Human kinase type associated with the inhibitor is shown in last column.

For common ligands, the increase in the number of protein structures, as well as ongoing development of structural (binding site) descriptors (e.g., [46]) is likely to allow for more widespread characterization of sites in the near future. Consequently, as in the case of established annotation tools, LigProf's predictive power will improve with PDB growth.

A number of approaches exist that allow for elucidation of conserved spatial motifs from structures without

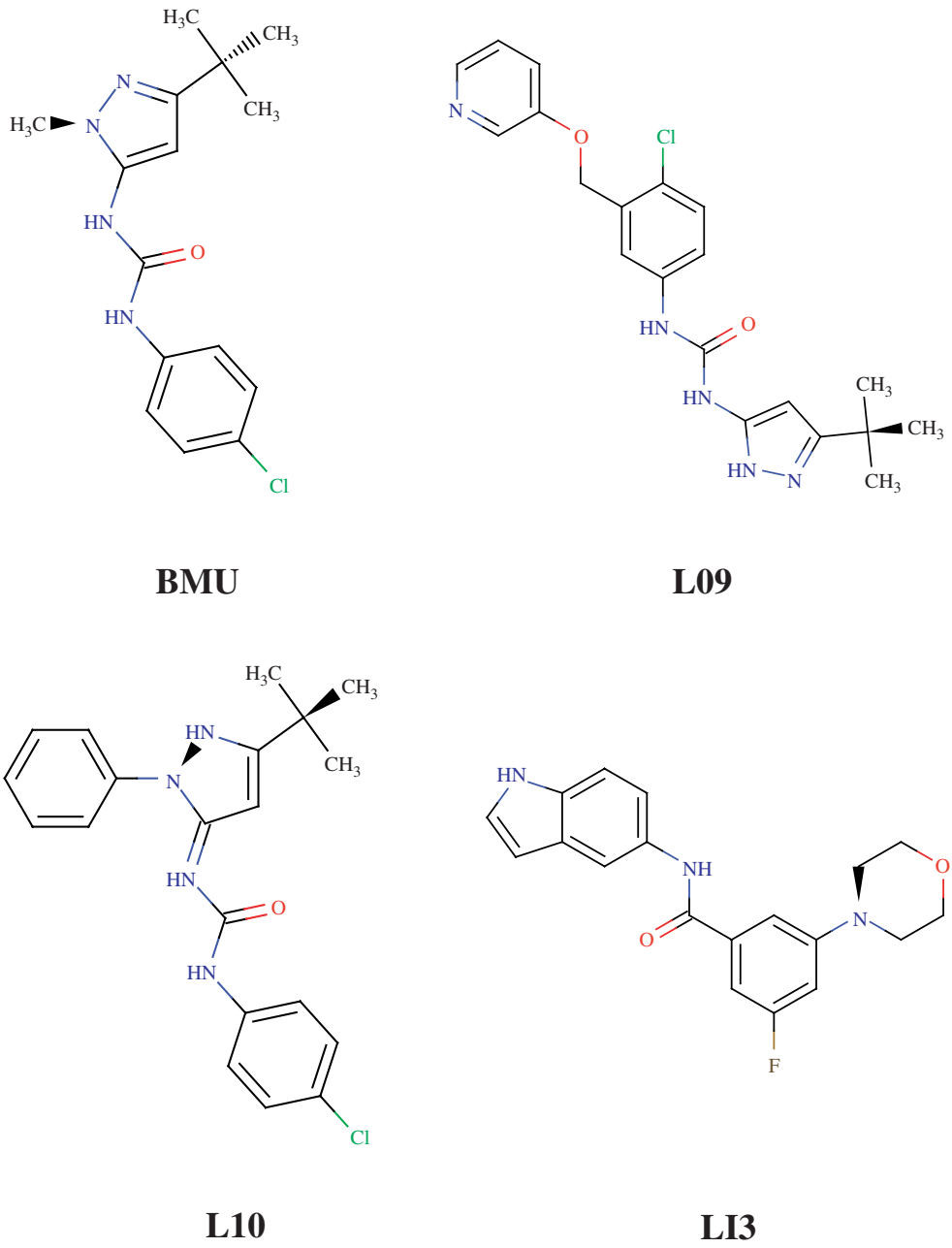


Fig. 4 Superimposition of identified potential inhibitors of US3 kinase (BMU, L09, L10, LI3), showing similar pharmacophore composition

bound ligands. This can be accomplished feasibly by matching (sub)structures, despite the assumed NP-hardness of the associated exact problem of clique detection (using a heuristic approach such as utilized by the SiteEngine [47]). Such geometric descriptors provide a means for future inclusion of non-complexed structures in a sequence-based approach. A further possibility lies in the use of scoring functions for physicochemical properties, such as the cavity detection approach utilized in the Ligsite algorithm [48] or Theoretical Microscopic Titration Curves, [49] both of which have already proven useful in elucidating functional sites from structures [50, 51].

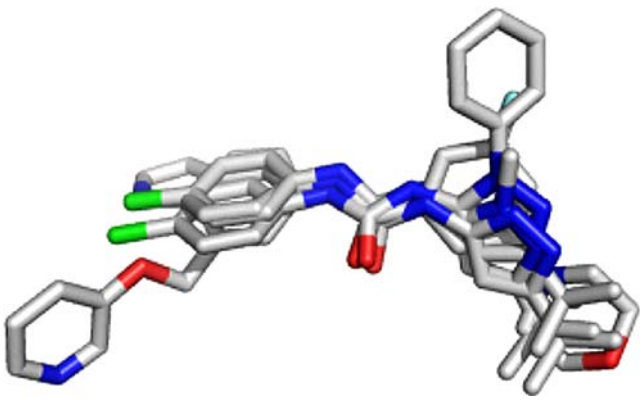


Fig. 5 Structures of identified potential inhibitors of US3 kinase (BMU, L09, L10, LI3)

A number of errors is likely to persist regardless of the coverage of different specificities in the reference database. There are presently three main sources of such persistent errors, all inherent to the PDB records being mined for binding sites [52].

The first and most common cause of misannotation is that the molecules cocrystallized with a protein are buffer molecules (or ions) without physiological significance (e.g., ethanediol, glycerol, phosphate anions). Second, the bound molecules are frequently sensitizers or transition-state analogues (e.g., the 545 and 485 compounds mentioned in the Results section). Such ligands share similar topology with the native substrate, but can have significant differences in chemical make-up. The differences stem from both different atoms making up the bulk of the molecule and adducts not present in the native substrate. While physiologically significant, the occurence of these ligand-binding sites requires expert interpretation. The third source of error is due to artificial binding sites formed during the co-crystallization process.

Addressing the above-mentioned sources of annotation errors, coupled with rapidly growing amounts of data provided by high-throughput structural genomics projects, marks batch, direct annotation of substrate specificity as a difficult but promising direction for future LigProf development.

Acknowledgements This work has been supported by grants from the following EU projects: DATAGENOM (LSHB-CT-2003-503017) and GENEFUN (LSHG-CT-2004-503567). LSW was supported by a Program for Young Researchers from the Foundation of Polish Science and MNiSW research grants (2 P05A 001 30, PBZ-MNiI-2/1/2005).

References

1. Hegyi H, Gerstein M (2001) Genome Res 11:1632–1640
2. Minshull J, Ness JE, Gustafsson C, Govindarajan S (2005) Curr Opin Chem Biol 9:202–209
3. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Nucl Acids Res 25:3389–3402
4. Devos D, Valencia A (2001) Trends Genet 17:429–431
5. Rost B, Liu J, Nair R, Wrzeszczynski KO, Ofran Y (2003) Cell Mol Life Sci 60:2637–2650
6. Green ML, Karp PD (2005) Nucl Acids Res 33:4035–4039
7. George RA, Spriggs RV, Thornton JM, Al-Lazikani B, Swindells MB (2004) Bioinformatics 20(Suppl 1):i130–i136
8. Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, Eddy SR, Sonnhammer EL, Bateman A (2006) Nucl Acids Res 34:D247–D251
9. Letunic I, Copley RR, Schmidt S, Ciccarelli FD, Doerks T, Schultz J, Ponting CP, Bork P (2006) Nucl Acids Res 32:D142–D144
10. Marchler-Bauer A, Anderson JB, Cherukuri PF, DeWeese-Scott C, Geer LY, Gwadz M, He S, Hurwitz DI, Jackson JD, Ke Z, Lanczycki C, Liebert CA, Liu C, Lu F, Marchler GH, Mullokandov M, Shoemaker BA, Simonyan V, Song JS, Thiessen PA, Yamashita RA, Yin JJ, Zhang D, Bryant SH (2005) Nucl Acids Res 33:D192–196
11. George RA, Spriggs RV, Bartlett GJ, Gutteridge A, MacArthur MW, Porter CT, Al-Lazikani B, Thornton JM, Swindells MB (2005) Proc Natl Acad Sci USA 102:12299–12304
12. Gold ND, Jackson RM (2006) J Mol Biol 355:1112–1124
13. Shatsky M, Shulman-Peleg A, Nussinov R, Wolfson HJ (2006) J Comput Biol 13:407–428
14. Stoll V, Stewart KD, Maring CJ, Muchmore S, Giranda V, Gu YG, Wang G, Chen Y, Sun M, Zhao C, Kennedy AL, Madigan DL, Xu Y, Saldivar A, Kati W, Laver G, Sowin T, Sham HL, Greer J, Kempf D (2003) Biochemistry 42:718–727
15. Terasaka T, Kinoshita T, Kuno M, Seki N, Tanaka K, Nakanishi I (2004) J Med Chem 47:3730–3743
16. Schafferhans A, Klebe G (2001) J Mol Biol 307:407–427
17. Carpy AJ, Marchand-Geneste N (2006) SAR QSAR Environ Res 17:1–10
18. Snyder KA, Feldman HJ, Dumontier M, Salama JJ, Hogue CW (2006) BMC Bioinformatics 7:152
19. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) Nucl Acids Res 28:235–242
20. Sobolev V, Sorokine A, Prilusky J, Abola EE, Edelman M (1999) Bioinformatics 15:327–332
21. Eddy SR (1998) Bioinformatics 14:755–763
22. Pils B, Copley RR, Schultz J (2005) BMC Bioinformatics 6:210
23. Sneath PHA, Sokal RR (1973) WH Freeman, San Francisco
24. Gribskov M, McLachlan AD, Eisenberg D (1987) Proc Natl Acad Sci USA 84:4355–4358
25. Tudos E, Cserzo M, Simon I (1990) Int J Pept Protein Res 36:236–239
26. Magliery TJ, Regan L (2005) BMC Bioinformatics 30:240
27. Lichtarge O, Bourne HR, Cohen FE (1996) J Mol Biol 257:342–358
28. Mihalek I, Res I, Lichtarge O (2004) J Mol Biol 336:1265–1282
29. Johnson JM, Church GM (2000) Proc Natl Acad Sci USA 97:3965–3970
30. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M (2003) Nucl Acids Res 31:365–370
31. Henikoff S, Henikoff JG (1992) Proc Natl Acad Sci USA 89:10915–10919
32. Wu TD, Nevill-Manning CG, Brutlag DL (2000) Bioinformatics 16:233–244
33. Sali A, Blundell TL (1993) J Mol Biol 234:779–815
34. Elofsson A (2002) Proteins 46:330–339
35. Kalinina OV, Mironov AA, Gelfand MS, Rakhmaninova AB (2004) Protein Sci 13:443–456
36. Hamelryck T, Manderick B (2003) Bioinformatics 19:2308–2310
37. Yang XL, Otero FJ, Skene RJ, McRee DE, Schimmel P, Pouplana L Ribas de (2003) Proc Natl Acad Sci USA 100:15376–15380
38. Li W, Jaroszewski L, Godzik A (2002) Bioinformatics 18:77–82
39. Qiu X, Janson CA, Smith WW, Green SM, McDevitt P, Johanson K, Carter P, Hibbs M, Lewis C, Chalker A, Fosberry A, Lalonde J, Berge J, Brown P, Houge-Frydrych CS, Jarvest RL (2001) Protein Sci 10:2008–2016
40. Hagglund R, Munger J, Poon AP, Roizman B (2002) J Virol 76:743–754
41. Kato A, Yamamoto M, Ohno T, Tanaka M, Sata T, Nishiyama Y, Kawaguchi Y (2006) J Virol 80:1476–1486
42. Prichard MN, Britt WJ, Daily SL, Hartline CB, Kern ER (2005) J Virol 79:15494–15502
43. Mestres J (2005) Drug Discov Today 10:1629–1637

44. Diwan P, Lacasse JJ, Schang LM (2004) J Virol 78:9352–9365
45. Schang LM, Coccaro E, Lacasse JJ (2005) Nucleosides Nucleotides Nucleic Acids 24:829–837
46. Guo T, Shi Y, Sun Z (2005) Prot Eng Des Sel 18:65–70
47. Shulman-Peleg A, Nussinov R, Wolfson HJ (2004) J Mol Biol 339:607–633
48. Schmitt S, Kuhn D, Klebe G (2002) J Mol Biol 323:387–406
49. Ondrechen MJ, Clifton JG, Ringe D (2001) Proc Natl Acad Sci USA 98:12473–12478
50. Kuhn D, Weskamp N, Schmitt S, Hullermeier E, Klebe G (2006) J Mol Biol 359:1023–1044
51. Ko J, Murga LF, Wei Y, Ondrechen MJ (2005) Bioinformatics 21: i258–i265
52. Kinoshita K, Nakamura H (2005) Protein Sci 14:711–718