VAGNER SOARES RIBEIRO

DETECÇÃO E VISUALIZAÇÃO DE SUBESTRUTURAS COMUNS NA INTERFACE PROTEÍNA-LIGANTE EM NÍVEL ATÔMICO ATRAVÉS DE MINERAÇÃO DE SUBGRAFOS FREQUENTES

Dissertação apresentada à Universidade Federal de Viçosa, como parte da exigências do Programa de Pós-Graduação em Ciência da Computação, para obtenção do título de *Magister Scientiae*

VIÇOSA
MINAS GERAIS - BRASIL
2018

Ficha catalográfica preparada pela Biblioteca Central da Universidade Federal de Viçosa - Câmpus Viçosa

T
R484d
2018

Ribeiro, Vagner Soares, 1984-
Detecção e visualização de subestruturas comuns na interface proteína-ligante no nível atômico através de mineração de subgrafos frequentes / Vagner Soares Ribeiro. – Viçosa, MG, 2018.
viii, 59f. : il. (algumas color.) ; 29 cm.

Inclui apêndices.
Orientador: Sabrina de Azevedo Silveira.
Dissertação (mestrado) - Universidade Federal de Viçosa.
Referências bibliográficas: f.34-37.

1. Mineração de dados (Computação). 2. Banco de dados. 3. Proteínas. 4. Bioinformática. I. Universidade Federal de Viçosa. Departamento de Informática. Programa de Pós-graduação em Ciência da Computação. II. Título.

CDD 22 ed. 005.74

VAGNER SOARES RIBEIRO

DETECÇÃO E VISUALIZAÇÃO DE SUBESTRUTURAS COMUNS NA INTERFACE PROTEÍNA-LIGANTE NO NÍVEL ATÔMICO ATRAVÉS DE MINERAÇÃO DE SUBGRAFOS FREQUENTES

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Ciência da Computação, para obtenção do título de *Magister Scientiae*.

APROVADA: 21 de fevereiro de 2018.

Adriana Maria Patarroyo Vargas

Jugurta Lisboa Filho

Sabrina de Azevedo Silveira
(Orientador)

Dedico especialmente à minha esposa Roberta e à minha filha Lia, pelo amor, paciência e carinho em todos os momentos dessa etapa em nossas vidas. Vocês estiveram sempre ao meu lado, mesmo nos momentos mais difíceis.
Dedico aos meus pais e ao meu irmão pelo amor e por sempre me incentivarem na busca pelos meus objetivos.

Agradecimentos

Agradeço especialmente à minha esposa Roberta e à minha filha Lia por todo apoio, amor, compreensão e paciência em todo este período.
Aos meus pais, por terem me criado com amor e atenção, e por sempre terem me incentivado em minhas escolhas.
Ao meu irmão Leandro, pela confiança e amizade.
À Universidade Federal de Viçosa, pela grande oportunidade de participar deste programa.
À inestimável ajuda de minha orientadora Sabrina, que auxiliou de forma ativa em todas as etapas de meu trabalho.
Ao Charles, Alexandre, Sócrates, Pedro e Samuel com quem trabalhei diretamente neste período.
À Raquel, Adriana, Valdete, Sandro e Maria Goreti que colocaram seus conhecimentos e experiência à disposição para auxiliar na pesquisa.
A todos os meus colegas de mestrado pela amizade e companherismo.
Aos professores do DPI pela paciência, comprometimento e amizade.
Ao Altino Alves pelo apoio durante este tempo.
A todos que contribuíram de diversas formas para que este momento fosse possível.

Sumário

Resumo

RIBEIRO, Vagner Soares, M.Sc., Universidade Federal de Viçosa, fevereiro de 2018. **Detecção e visualização de subestruturas comuns na interface proteína-ligante no nível atômico através de mineração de subgrafos frequentes**. Orientadora: Sabrina de Azevedo Silveira

O volume de dados disponíveis sobre sequências e estruturas de proteínas vem crescendo rapidamente ao longo dos últimos anos. Por exemplo, o banco de dados de sequências Uniprot possui hoje mais de 500.000 sequências revisadas manualmente e outras 107 milhões automaticamente anotadas. A quantidade de estruturas disponíveis é menor, mas ainda assim bem extensa, onde temos no PDB (*Protein Data Bank*), principal repositório de estruturas proteicas, mais de 137 mil estruturas depositadas. Neste contexto, a Bioinformática desempenha um importante papel, no sentido de armazenar, organizar e extrair informação da grande quantidade de dados provenientes de organismos vivos. Neste trabalho, são propostas duas abordagens visuais para dados em larga escala que auxiliem os pesquisadores a responder diferentes questões relacionadas à proteínas. No primeiro trabalho, VERMONT (*ViewER MutatiON Tool*) 2.0, é apresentada uma plataforma interativa que utiliza dados de sequência e estruturas de proteínas, com o objetivo de fornecer informações sobre pontos de mutação potencialmente danosos à estrutura e função proteica. Uma mutação é uma alteração na sequência de DNA, podendo afetar um único par de bases ou múltiplos genes. Elas podem ser hereditárias, quando herdadas dos pais ou somáticas, quando ocorrem durante a vida. O VERMONT tem o objetivo de permitir o estudo do efeito de pontos de mutação em

proteínas, que ocorrem quando um único nucleotídeo é alterado na sequência de DNA, podendo modificar determinados resíduos na proteína . Para isto, é utilizada uma abordagem mista, que envolve interações intramoleculares, cálculo de acessibilidade a solvente, métricas de redes complexas e cálculo da energia livre de Gibbs. O cálculo destas informações acoplado a uma amigável interface, permite aos especialistas analisar e compreender o impacto de pontos de mutações. O segundo trabalho, visGReMLIN (*visual Graph Mining strategy to infer protein-Ligand INteraction patterns*), utiliza uma abordagem de mineração de grafos para encontrar padrões de interação recorrentes entre proteínas e ligantes. São considerados ligantes pequenas moléculas não proteicas que formam um complexo com a proteína. Nele, as interações entre proteína e ligante são modelados como um grafo bipartido, onde os vértices são rotulados com as características físico-químicos do átomo e as arestas com o tipo de interação estabelecido pelos átomos. Para encontrar os padrões de interação, foi realizado um agrupamento dos grafos, utilizando, por padrão, o método k-medoids com coeficiente de silhueta como métrica de avaliação, seguido por uma mineração de subgrafos frequentes. Esta estratégia foi capaz de identificar átomos e resíduos importantes, já determinados experimentalmente ou computacionalmente, em estudos anteriores. Em ambos os trabalhos, foram propostas ferramentas robustas e fáceis de usar que lidam com dados de proteínas em larga escala e tem o objetivo de auxiliar na compreensão de diferentes problemas biológicos.

Abstract

RIBEIRO, Vagner Soares, M.Sc., Universidade Federal de Viçosa, February, 2018. **Graph mining based detection and visualization of conserved motifs in 3D protein-ligand interface at atomic level**. Adviser: Sabrina de Azevedo Silveira

Protein sequences and structures is growing fast in the last years. For example, the database of protein sequences Uniprot has more than 500.000 sequences manually annotated and others 107.000.000 automatically annotated. The number of protein structures available is smaller, but still very extensive, with 137000 structures deposited in the PDB (Protein Data Bank), major protein structure repository. In this context, bioinformatics plays an important role, when allows to store, organize and deal with the huge amount of data arising from living organisms. In this work, two visual approaches are proposed to handle large scale data to help researchers to answer different questions about proteins. In the first work, VERMONT (ViewER MutatiON Tool) 2.0, it's presented an interactive plataform that uses protein sequence and structure data to provide information about point mutations potentially harmful to the protein structure or function. To do so, different approaches are used as intramolecular interactions, solvent accessibility calculation, complex network metrics, e Gibbs free energy. The computation of this information coupled with a friendly interface enables specialists to analyse and make sense of the impact of point mutations. Mutation is the alteration in the DNA sequence that may affect a single base pair or multiple genes. They may be hereditary, when inherited from parents or somatic, when occur during the life. The VERMONT

aim is to allow the study of the impact of point mutations in proteins, that happens when a single nucleotide is modified in the DNA sequence, what can alter protein residues. The second work, visGReMLIN (visual Graph Mining strategy to infer protein-Ligand INteraction patterns), uses a graph mining approach to detect motifs in the protein-ligand interaction. Are considered as ligands, small non protein molecules that forms a complex with the protein. The interactions between protein and ligand are modeled as a bipartite graph, where nodes are labeled with the physicochemical properties of atoms and the edges are labeled with the interaction type established between atoms. To detect the motifs, the graphs are clustered, using, by default, the k-medoids method and the silhouette coefficient as evaluation metric, followed by a frequent subgraph mining. This strategy was able to find relevant atoms and residues experimentally or computationally determined in previous studies. In both works, are provided robust and easy to use tools that deal with large scale protein data and aim to help in the understanding of different biological problems.

Capítulo 1

Introdução

A compreensão de fenômenos biológicos é de grande importância para diversas aplicações como descoberta de medicamentos, controle de pragas, técnicas terapêuticas individualizadas, dentre outras. Neste sentido, a Bioinformática desempenha um papel muito importante, por permitir lidar com a grande quantidade de dados proveniente do estudo de organismos vivos, como sequências de DNA, estrutura de proteínas, caminhos metabólicos, etc. É um campo acadêmico que busca criar e aprimorar algoritmos, técnicas computacionais, estatísticas, e teoria para resolver problemas práticos e formais que surgem a partir da análise de dados biológicos. Ela se concentra primordialmente em problemas que envolvam dados que emergem de células de organismos vivos (Nair, 2007).

Em linhas gerais são três os objetivos da bioinformática: (i)Organizar os dados de uma forma que possibilite o acesso dos pesquisadores à informação existente. (ii)Desenvolver ferramentas que auxiliem na análise dos dados,o que vai além de meras buscas textuais, pois é necessário compreender o significado biológico dos padrões encontrados (iii)Usar as ferramentas construídas para analisar e interpretar os resultados do ponto de vista biológico (Luscombe et al., 2001).

Este trabalho envolveu duas diferentes frentes de atuação, que resultaram em modelos visuais(VERMONT e visGReMLIN) para que pesquisadores possam extrair informações relevantes que os possibilite responder questões biológicas relacionadas a proteínas. Proteínas são uma classe extremamente importante de macromoléculas, por desempenharem diferentes papéis nas células, como a cons-

trução de tecidos humanos, o transporte de substâncias, como o oxigênio, a defesa do organismo através de anticorpos, a catalisação de reações químicas através das enzimas, a regulação de hormônios, dentre outras. Essencialmente, as proteínas desempenham suas funções através da interação física com outras moléculas, como outras proteínas, peptídeos, ácidos nucléicos, substratos, como oxigênio, solvente e metal (Xing et al., 2016). Nas últimas décadas, muitas estratégias computacionais foram propostas para medir a influência das mutações na estrutura e função das proteínas. Como exemplo, podemos citar o SDM (Site Directed Mutator) (Topham et al., 1997), NeEMO (Giollo et al., 2014), MAESTRO (Laimer et al., 2015), iStable (Chen et al., 2013) e DUET (Pires et al., 2014). Apesar disso, cada estratégia por si só não se mostrou precisa para todos os cenários de mutação (Pires et al., 2014). Devido a esta fato, a estratégia de combinar diferentes métodos vêm ganhando atenção, como é feito pelo iStable e DUET. Neste sentido, é apresentado o primeiro trabalho, VERMONT 2.0 (*ViewER MutatiON Tool*), como descrito no capítulo 2 . Ele aprimorou o VERMONT (SILVEIRA et al., 2014), apresentado no *Biovis Contest* 2013 e premiado com o *Biology Experts Pick*, que tinha o objetivo de encontrar pontos de mutação que poderiam impactar na função da proteína e como ela poderia ser restaurada. Os dados usados foram fornecidos pelos organizadores da competição, sendo constituídos da triosefosfato isomerase defeituosa (dTIM) e a *S. cerevisiae* (scTIM) baseado num conjunto de dados de proteínas da mesma família. A sua limitação reside no fato de permitir que apenas um conjunto de dados específico seja analisado. Por este motivo, o VERMONT 2.0 foi completamente reimplementado, permitindo agora receber como entrada qualquer estrutura de proteína no formato do PDB (*Protein Data Bank*). O usuário pode buscar por sequências com um determinado percentual de similaridade, ou informar suas próprias estruturas de interesse. Como primeiro passo, o VERMONT 2.0 utiliza o Multiprot (Shatsky et al., 2014) para realizar um alinhamento de sequências baseado em estrutura par a par da proteína selvagem com cada uma das proteínas da família. Com base neste alinhamento, a estratégia de visualização permite que especialistas visualizem o alinhamento, as interações intramoleculares, propriedades físico-químicas e parâmetros topológicos de redes complexas. Cada um dos diferentes modos de visualização é descrito a seguir.

Módulo de alinhamento de sequência baseado em estrutura: Neste módulo, cada linha representa uma cadeia proteica, com a sequência da proteína mutante no topo, logo acima da proteína selvagem. Para facilitar a visualização de tendências e exceções, linhas similares são agrupadas através do algoritmo EM (Expectation Maximization). Cada coluna representa uma posição do alinhamento. Para auxiliar na identificação de posições conservadas, 3 esquemas de cores são disponibilizados (CINEMA, CLUSTAL e LESK). Com isto os resíduos de cada cadeia são coloridos de acordo com suas propriedades físico-químicas.

Módulo de propriedades físico-químicas: A acessibilidade ao solvente é calculada, usando o software Naccess que implementa o algoritmo de Lee e Richards (LEE, 1971). Os resultados são exibidos através de um mapa de calor, para cada acessibilidade calculada pelo Naccess. Para cada resíduo, uma intensidade de cor é associada. Quanto maior for a acessibilidade, maior será a intensidade da cor. Isto auxilia na identificação visual de posições no alinhamento com grau de acessibilidade conservadas.

Módulo de interações: As interações intramoleculares de cada cadeia proteica são representadas como um grafo, onde os vértices representam átomos e as arestas correspondem às interações entre eles. As interações são calculadas usando triangulação de Delaunay (Okabe et al., 2008), através da biblioteca CGAL (CGAL, 2017). Esta estratégia tem o objetivo de excluir interações obstruídas. Os vértices são rotulados de acordo com suas características físico-químicas, como descrito por Gonçalves-Almeida (Gonçalves-Almeida et al., 2011). Arestas são rotuladas de acordo com os tipos dos átomos envolvidos na interação e um critério de distância (Mancini et al., 2004). As interações então são mapeadas a nível de resíduo. A estratégia de visualização usada neste módulo permite visualizar as interações de determinado tipo, permitindo identificar as posições conservadas ao longo da família de proteínas. Além disso, é possível selecionar um determinado resíduo e visualizar suas interações em uma representação molecular tridimensional ou, bidimensional, através de um grafo.

Módulo de propriedades topológicas: São computadas 3 métricas de centralidade de redes complexas para auxiliar na identificação de vértices potencialmente importantes na rede de interações de cada cadeia proteica. O primeiro deles é o Degree (grau), que representa o número de arestas conectados a determi-

nado vértice. O segundo é o Betweenness, que mede o quanto um vértice está no menor caminho entre dois outros vértices quaisquer da rede. Por último, temos o Closeness, que é a distância média de um vértice para todos os outros. A visualização de mapa de calor para este módulo tem o objetivo de auxiliar na identificação de resíduos importantes a partir de uma perspectiva de redes complexas.

Além do apresentado anteriormente, o VERMONT 2.0 ainda permite exibir apenas colunas da proteína mutante ou de sítios catalíticos, usando como referência o CSA (Catalytic Site Atlas) (Furnham et al., 2014) , filtrar posições do alinhamento por percentual de conservação e avaliar os efeitos causados por uma mutação através da variação da energia livre de Gibbs.

Esta estratégia integrada permitiu explorar com sucesso pontos de mutações analisados experimentalmente para o núcleo da proteína supressora de tumor, p53, responsável por muitas mutações que levam ao câncer em humanos. Outros pontos de mutação potencialmente danosos também foram sugeridos. Este exemplo demonstra o potencial do VERMONT em auxiliar especialistas na obtenção de informações relevantes para a identificação de mutações impactantes de forma integrada e totalmente visual.

No capítulo 3, é apresentado o visGReMLIN que trata de interações entre proteínas e ligantes.

As interações entre proteínas e ligantes são fundamentais para uma grande variedade de processos que ocorrem em organismos vivos. Pequenas moléculas ligantes, como metabólitos ou medicamentos, constantemente interagem com seus respectivos receptores proteicos (Gao & Skolnick, 2013b). A evolução da função de uma proteína é dependente, em parte, do desenvolvimento de locais altamente específicos, projetados para vincular pequenas moléculas ligantes com afinidades ajustadas às necessidades da célula (Dunn, 2007). A compreensão destas interações tem sido objeto de estudo de vários trabalhos de pesquisa (Gao & Skolnick, 2013a) (Kadukova & Grudinin, 2017) (Kadukova & Grudinin, 2017) (Pai et al., 2017) (Chandel et al., 2017), uma vez que a identificação de interações conservadas entre proteínas e ligantes é primordial no processo de reconhecimento molecular e pode contribuir para o desenvolvimento racional de drogas, predição de ligantes e identificação de alvos biológicos. Segundo Tuncbag (Tuncbag et al., 2011), estruturas proteicas são mais conservadas que

suas sequências, e resíduos formadores de interface (RFI) são ainda mais conservados. RFIs são regiões de interface molecular entre proteínas. No nosso trabalho, focamos na interface entre proteínas e ligantes. Neste caso, consideramos ligantes como pequenas moléculas não proteicas.

Vários métodos já foram propostos para identificar padrões de interação tridimensionais. Como exemplo, temos o método desenvolvido por Kuttner (Kuttner et al., 2003) e Nebel (Nebel et al., 2007), que envolve a sobreposição de proteínas baseados em um ligante, seguido por um agrupamento de resíduos e átomos conservados. Gonçalves-Almeida (Gonçalves-Almeida et al., 2011), desenvolveu um método baseado em centróides hidrofóbicos, onde as RFIs são modeladas como grafos. Grafos também foram usados por Pires (Pires et al., 2013), para gerar uma assinatura que compreende padrões de distância para pockets de proteínas. Nenhum dos métodos citados, entretanto, consideram ligantes. Desaphy (Desaphy et al., 2013), por sua vez, criou um fingerprint genérico usando informação estrutural da interação proteína-ligante codificada em grafos. Por último, cito o LibME (He et al., 2016) que extrai padrões tridimensionais de ligação usando informações do ligante e dos resíduos que estão ao seu redor.

Neste trabalho, o visGReMLIN (visual Graph Mining strategy to infer protein-Ligand INteraction patterns), foi desenvolvida uma plataforma visual interativa que implementa o GReMLIN (Santana et al., 2016), que foi completamente reescrito e aprimorado. Nossa estratégia é baseada em mineração de grafos e tem como objetivo detectar padrões de interação conservados em larga escala para um conjunto de interações proteína-ligante. Para isto, a interface proteína-ligante foi modelada como um grafo bipartido, onde os vértices representam átomos e as arestas correspondem às interações entre os átomos. Cada vértice é rotulado de acordo com suas propriedades físico-químicas, e as arestas de acordo com os tipos de átomos envolvidos na interação e a distância entre eles. Os grafos resultantes correspondem às interações não covalentes entre cada cadeia proteica do conjunto de dados e seus respectivos ligantes. Com estes grafos em mãos, é realizado um agrupamento seguido de uma mineração de subgrafos frequentes para cada um dos grupos encontrados.

A plataforma visual permite ao usuário buscar por cadeias de proteínas similares a uma determinada cadeia de referência disponível no PDB, informar suas

próprias cadeias de interesse ou realizar o upload de estruturas ainda não depositadas no PDB. Além disso, é permitido ao usuário alterar alguns parâmetros de configuração, como ,por exemplo, os limiares de distância para cada tipo de interação e o algoritmo de agrupamento. Após o GReMLIN ser executado, vários módulos são fornecidos para que os pesquisadores possam explorar os resultados encontrados.

Módulo de detalhes do conjunto de dados: Permite explorar os resultados do agrupamento realizado nos grafos. Nele é possível visualizar os grupos encontrados e os grafos que pertencem a cada grupo, assim como os ligantes encontrados em determinado grupo.

Módulo de tabela de padrões: Sumariza os padrões encontrados para cada grupo e valor de suporte.

Módulo de visualização de padrões: Possibilita ao usuário explorar os grafos de interação encontrados em duas ou três dimensões e aplicar os diversos filtros disponíveis para obter informações de seu interesse. Além disso, permite relacionar os padrões aos grafos onde eles foram encontrados. Este relacionamento é feito através do algoritmo VF2 (Cordella et al., 2004).

Módulo de análise gráfica: Exibe gráficos de contagem para cada tipo de átomo e interação encontrados nos padrões.

O visGReMLIN foi testado para dois conjuntos de dados com características diferentes. No primeiro deles, foram usadas 73 cadeias da proteína CDK (Cyclin-dependent kinases) em complexo com diferentes ligantes, para demonstrar um caso em que temos promiscuidade da proteína. No segundo, foram utilizadas 50 cadeias em complexo com o ligante ATP (Adenosine triphosphate) para analisar uma situação com promiscuidade de ligante. Em ambos os casos, foi possível encontrar átomos/resíduos importantes já determinados experimentalmente ou computacionalmente, de maneira visual e interativa, demonstrando um importante resultado do trabalho.

No VERMONT auxiliei na implementação de alguns algoritmos, enquanto no visGReMLIN atuei na generalização do GReMLIN e em alguns recursos de visualização. Várias tecnologias foram utilizadas neste trabalho como python, perl, C++, shell script e PHP no lado do servidor, além de javascript no lado do

cliente, apoiado por frameworks como D3 e 3Dmol, utilizados em diversos recursos de visualização.

O capítulo 2 detalha o VERMONT, através de um artigo científico. No apêndice A está disponível o material suplementar deste trabalho.

O capítulo 3 descreve o visGReMLIN por meio de um artigo científico. O material suplementar pode ser encontrado no Apêndice B.

Capítulo 2

Vermont: a multi-perspective visual interactive platform for mutational analysis

Artigo científico publicado na BMC Bioinformatics.

Capítulo 3

Detecção e visualização de subestruturas comuns na interface proteína-ligante no nível atômico através de mineração de subgrafos frequentes

Artigo científico submetido à conferência ISMB (*International Conference on Intelligent Systems for Molecular Biology*) 2018.

Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

OXFORD

Subject Section

Detecção e visualização de subestruturas comuns na interface proteína-ligante em nível atômico através de mineração de subgrafos frequentes

Vagner S. Ribeiro [1]*, Charles A. Santana [2], Alexandre V. Fassio [2], Adriana M. Patarroyo-Vargas [3], Maria G. A. Oliveira [3, 4], Valdete M. Gonçalves-Almeida [2], Sandro C. Izidoro [5] , Raquel C. de Melo-Minardi [6] and Sabrina de A. Silveira [1,*]

[1]Department of Computer Science, Universidade Federal de Viçosa, Viçosa, 36570-900, Brazil and
[2]Department of Biochemistry and Immunology, Universidade Federal de Minas Gerais, Belo Horizonte, 31270-901, Brazil and
[3]Department of Biochemistry and Molecular Biology, Universidade Federal de Viçosa, Viçosa, 36570-900, Brazil and
[4]Instituto de Biotecnologia aplicada à Agropecuária (BIOAGRO), Universidade Federal de Viçosa, Viçosa, 36570-900, Brazil and
[5]Department of Computer Engineering, Advanced Campus at Itabira, Universidade Federal de Itajubá, Itabira, 35903-087, Brazil and
[6]Department of Computer Science, Universidade Federal de Minas Gerais, Belo Horizonte, 31270-901, Brazil and

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Interactions between proteins and non-proteic small molecule ligands play important roles in biological processes of living systems. Thus the development of computational methods to support on the understanding of the ligand-receptor recognition process is of fundamental importance, since this this is a major step towards ligand prediction, target identification, lead discovery, among others. This article presents visGReMLIN, a web server that couples a graph mining based strategy to detect motifs in the protein ligand interface and an interactive platform to visually explore and interpret such motifs in the context of protein-ligand interfaces.
Results: We intended to contribute a visual analytics oriented web server to detect and visualize common motifs in protein-ligand interface. To illustrate the ability of visGReMLIN to do so, we conducted 2 use cases in which our strategy was compared with previous experimentally and computationally determined results. Our strategy allowed us to detect patterns previously documented in the literature in a totally visual manner. In addition, we found some motifs we believe are relevant for protein-ligand interaction in the analyzed datasets.
Availability: http://200.235.159.97/visgremlin3
Contact: sabrina@ufv.br
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

At the molecular level, protein receptors constantly interact with small-molecule ligands such as metabolites or drugs. A variety of protein functions can be attributed to or regulated by these interactions (Gao and Skolnick, 2013b). Understanding how protein-ligand interactions take place has been the goal of many research works (Gao and Skolnick, 2013a; Kadukova and Grudinin, 2017; Pai *et al.*, 2017; Chandel *et al.*, 2018) as molecular recognition is pivotal in biological processes in living organisms, including signal transduction, catalysis and regulation of biological function, to name a few.

Identifying conserved interactions between proteins and ligands that are reused across a protein family is a key factor for understanding molecular recognition processes and can contribute to rational drug design, target identification, lead discovery and ligand prediction. Interface

forming residues (IFR) are those in regions of molecular interface between proteins. In accordance with Tuncbag *et al.* (2011), protein structures are more conserved than their sequences and IFRs are even more conserved than the whole protein structure. Therefore, IFR can be an invaluable source of information to support on the identification of conserved interactions across a set of complexes. In this paper we are interested in the interface between proteins and ligands. We consider as ligands small non protein molecules. On one hand, proteins can be promiscuous, as they interact with different ligands (Nobeli *et al.*, 2009; Kufareva *et al.*, 2011). On the other hand, ligands can also be promiscuous, when one ligand is recognized by different proteins (Cobanoglu *et al.*, 2013). Thus it is reasonable to expect that methods to detect conserved interactions between proteins and ligands should be able to deal with both protein and ligand promiscuity.

Several methods have been proposed to identify three-dimensional binding motifs. Here we briefly review some recent works that are representative examples of the diverse techniques already proposed.

Previous solutions for detecting structural binding motifs for a set of diverse proteins and a common ligand involved the superimposition of proteins based on the ligand, and the subsequent clustering of conserved residues or atoms interacting with such ligand. The methods developed by Kuttner *et al.* (2003) and Nebel *et al.* (2007) are examples of these kind of solutions. These strategies work well for rigid ligands as they result in structural alignments of good quality due to the ligand induced superimposition. In a more general manner, classical methods as sequence/structural alignment are not appropriate to detect conservation when proteins have dissimilar sequences and/or structures (Bonham-Carter *et al.*, 2013; Wang *et al.*, 2013; Vinga, 2014).

Gonçalves-Almeida *et al.* (2011) developed a method based on hydrophobic patch centroids to predict which they named cross-inhibition, also known as inhibitor promiscuity, in serine proteases. IFRs were modeled as a graph in which hydrophobic atoms are the nodes and the contacts between them are the edges. Centroids were used to summarize connected components of this graph and conserved centroids, termed hydrophobic patches, were used to characterize, detect and predict cross-inhibition.

In a similar manner, Pires *et al.* (2013) used graphs that consider physicochemical properties of atoms and their contacts to represent protein pockets, generating a signature that perceives distance patterns from protein pockets. Each binding site is represented by a feature vector that encodes a cumulative edge count of contact graphs defined for different cut-off distances, which are used as input data for learning algorithms. This signature does not require any ligand information and also it is independent of molecular orientation.

The motifs computed by the above methods, designed by Gonçalves-Almeida *et al.* (2011) and Pires *et al.* (2013), are able to identify, compare, classify and even predict binding sites. However, these motifs include only information on the protein side, and do not represent the non-covalent interactions established between the ligand and the receptor.

Desaphy *et al.* (2013) encoded structural protein-ligand interaction in graphs and simplified this information in a generic fingerprint, which is a vector of 210 integers, to encompass protein-ligand interaction patterns. To generate the fingerprint, each interaction is described by a pseudoatom. Then all possible pseudoatom triplets are counted within six distance ranges. Finally, the full vector is pruned to keep the most frequent triplets, resulting in the definition of a frame-invariant fingerprint. In addition to the fingerprint, authors developed two computational methods to align protein-ligand complexes from their interaction patterns.

Nakadai *et al.* (2016), in turn, introduced a method, based on the differences between residues that were superimposed on small molecule inhibitors and those that were not superimposed, for identifying key residue pairs as potential targets of protein-protein interfaces. This method deals with a set of complexes composed by similar proteins in complex with different ligands and can support on the rational design of inhibitors that target these interfaces. To classify which are the superimposed residues, authors performed structural alignments.

Recently, a method has been proposed to extract three dimensional binding motifs that capture information of the ligand and its surrounding residues in protein-ligand complexes. Based on the observation that the molecular function of a protein is frequently carried out through a limited number of residues, which are reused in functional conserved proteins during evolution, LibME (He *et al.*, 2016) searches for pocket residues, conserved in terms of chemical property and spatial position, that are situated around the target ligand. Thus the resulting motifs are composed only of residues. Also, the computation of pocket residues positions relative to the ligand avoids the ligand-induced alignment of pockets. This method is specific for diverse proteins binding the same or similar ligands.

In this paper we propose visGReMLIN, (visual Graph Mining strategy to infer protein-Ligand INteraction patterns), a user-friendly web server implementation of our method GReMLIN (Santana *et al.*, 2016) that uses a graph mining based strategy to detect conserved structural motifs in large scale datasets of protein-ligand interactions. visGReMLIN provides a visual interactive platform to support domain specialists on the detection of trends and exceptions in protein-ligand interactions allowing them to explore and make sense of the motifs.

In order to detect common substructures, here called patterns or motifs, in protein-ligand interface, we devised a graph mining based strategy which models the interface as bipartite graphs where nodes represent atoms from protein or from ligand and edges represent interactions between atoms. Here we are interested in non covalent interactions. Nodes are labeled according to their physicochemical properties, and edges are labeled according to interatomic interactions and distance criteria. Next, we perform a clustering analysis on these graphs, followed by a frequent subgraph mining on each cluster to detect motifs in protein-ligand interface.

In addition, we propose a visual interactive platform to explore protein-ligand interactions and their motifs. The input module automatically searches the Protein Data Bank (PDB) (Berman *et al.*, 2000) for similar structures given an entry informed by the user and a similarity threshold, or the user can enter a list of PDB entries. Also, visGReMLIN allows users to inform their own structural files (those that are not yet deposited in PDB). visGReMLIN then proceeds to the computations and notifies the user when the analysis is complete. In our visual representations of protein-ligand interfaces, we use color as a pre-attentive atributte that encodes physicochemical properties of atoms (nodes) and interactions (edges). Hence, users can see at a glance general trends and exceptions regarding properties of atoms and interactions. Moreover, we provide a variety of filters to explore interactions and their motifs as well as a text search to help users to find residues/atoms in which they are particularly interested. By typing the residue/atom in the text box, our tool highlights the corresponding nodes in the visualization. To support domain specialists on the understanding of motifs, visGReMLIN allows to select a specific motif (frequent subgraph) and highlights it in the context of interface graphs. These graphs are depicted as 2D schematic representation as well as in a 3D molecular viewer.

visGReMLIN is a large-scale, alignment-free strategy and results do not depend on molecular orientation. Furthermore, our method is not specific for a dataset of different proteins with the same/similar ligands or a dataset with the same/similar proteins and different ligands, being able to be used in both types of datasets. It means that visGReMLIN can be used with datsets of promiscuous proteins or ligands.

GReMLIN was a work in progress when presented at IEEE BIBE 2016. Although our strategy seemed promising, as it received the *Distinguished*

Student Paper Award[1] at that conference, it was a stand alone tool coupled with a prototype visualization able to deal with only 2 static predefined datasets. We implemented a whole new GReMLIN from scratch, improving unsupervised learning, frequent subgraph mining and the feature vectors that represent protein-ligand interactions. In addition, the visualization tool was also completely reimplemented and several new features were included, for instance: (i) visGReMLIN takes as input any protein-ligand complex in PDB format; (ii) each user can store projects executed in his/her account, facilitating the reproducibility of the experiment; and (iii) a user can share a project, allowing multiple users to view and explore the same project, which makes easier the collaboration among domain specialists.

2 Methods

In this section, we explain visGReMLIN based on the information flow in our tool. In addition, we detail the experiments design and datasets used to test and evaluate our strategy.

Before starting using visGReMLIN, the user needs to register and later login. This is useful to organize the projects in a same place, with no need to bookmark many different submissions. Beyond this, the user will receive the results via email once the process finishes.

Figure 1 shows the visGReMLIN workflow. The web-server has 3 main building blocks, which are *Creating a project*, *GReMLIN strategy* and *Data analytics visualization*. Next, we refer to Figure 1 in order to describe each step performed by visGReMLIN.

2.1 Creating a project

When starting a new project in visGReMLIN, there are 3 options for the user to provide the dataset of protein-ligand complexes (by complex we mean a PDB id and chain) to be analyzed (Figure 1-A):

- Inform a PDB id and chain and let our tool to automatically search on PDB for these complexes, by selecting an alignment method and an identity percentage;
- Enter a dataset of previously selected complexes manually (type or copy and paste);
- Upload user's own complexes in PDB format (structures that are not deposited in PDB).

visGReMLIN *Input module* is shown in Figure 1 from Supplementary Material. In addition to selecting the input dataset, users can choose the cutoff they want to use for interaction computation (see Section 2.2.1), the clustering algorithm and their evaluation metric (see Section 2.2.2) in *Input module*.

2.2 GReMLIN strategy

We use the term GReMLIN to refer to our large-scale, graph-based strategy to detect motifs in the protein-ligand interface. GReMLIN is composed of three main building blocks, which we detail next.

2.2.1 Graph dataset generation

The first step of GReMLIN is to retrieve from PDB the input dataset informed in the *Creating a project*, which is composed of set of protein-ligand complexes (Figure 1-B).

Then we proceed to hydrogen and experimental artifacts removal. In PDB, there are structures solved using different experimental methods, for instance, x-ray crystallography and nuclear magnetic resonance (NMR). Although Woińska *et al.* (2016) have recently stated that hydrogen atoms can be located by x-ray crystallography, PDB files from structures solved by NMR contain positions of hydrogens while those from structures solved by x-ray crystallography do not. Thus, to deal with structures obtained by different experimental methods in a fair manner, we removed hydrogens from PDB complexes before contact computation. Ligands with less than 6 atoms are considered experimental artifacts to solve protein structure and are removed. We keep only those complexes that contain at least one ligand.

Table 1. Distance criteria (in Å) and physicochemical types of atoms involved in each type of interaction

Type of interaction	Atom types	Minimum distance	Maximum distance
Aromatic stacking	two aromatic atoms	1.5	3.5
Hydrogen bond	an acceptor and a donor atom	2.0	3.0
Hydrophobic	two hydrophobic atoms	2.0	3.8
Repulsive	two atoms with the same charge	2.0	6.0
Salt bridges	two atoms with opposite charge	2.0	6.0

To compute protein-ligand interactions (PLI), we use a cutoff dependent approach coupled with a distance criteria. According to previous works (da Silveira *et al.*, 2009; Pires *et al.*, 2011), a contact is defined between a pair of atoms, (i, j), if the Euclidean distance between them is less than or equal to a cutoff distance. Thus we perform contact computation between protein and ligand atoms and, based on the physicochemical types of atoms and the distance between them, we define the type of interaction established, in a similar manner to Gonçalves-Almeida *et al.* (2011); Silveira *et al.* (2014); Santana *et al.* (2016); Fassio *et al.* (2017). Table 1 from Supplementary Material presents a list with atom types used in visGReMLIN, which was derived from Fassio *et al.* (2018). Table 1 provides the standard distance criteria and physicochemical types of atoms involved in each type of interaction. Physicochemical types of atoms from ligands were computed using Pmapper from Chemaxon[2] at pH 7. It is important to point out that the user is allowed to select his/her own distance criteria, as shown in Figure 1 from Supplementary Material.

In a bipartite graph $G = (P, L, E)$, nodes can be divided into two disjoint sets, P and L , such that every edge in E connects a vertex in P to one in L (Diestel, 2000). We model PLI as bipartite graphs where nodes depict atoms from protein (P) or ligand (L) and edges represent interactions (E) among them. Nodes and edges are labeled according to the physicochemical types of atoms and interactions. Figure 2 shows an exemple of PLI represented by a bipartite graph. The block *Graph dataset generation* has these bipartite graphs as output.

2.2.2 Unsupervised learning

This block (Figure 1-C) takes as input a set of graphs that represent the interfaces between proteins and ligands and segments them in similar groups through an unsupervised learning strategy for motif prediction in the next block.

To summarize our dataset of graphs, we modeled the dataset of PLI graphs as a matrix that contains information from nodes and edges properties. For each graph, we generate a feature vector in which each position represents the presence of specific properties on a certain graph. Each vector position represents a pair of node properties separated by the edge distance (number of edges between the pair of nodes). To calculate such feature vector, we performed a breadth first search on each graph to obtain each non cyclic path between all pairs of nodes. These paths are represented by their end nodes with their respective node properties as

[1] http://bibe2016.asia.edu.tw/announcement/

[2] http://www.chemaxon.com

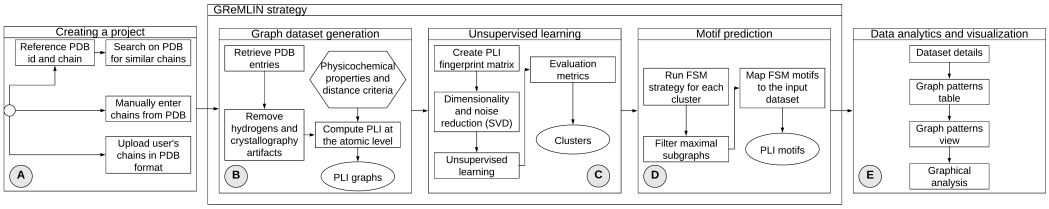


Fig. 1. visGReMLIN workflow. The workflow is divided in three main building blocks: Creating a project; GReMLIN strategy ; Data analytics and visualization. The circle represents the starting point; rectangles denote processing steps; ellipsoids represent output files; hexagons are input (files or parameters).

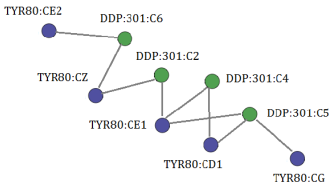
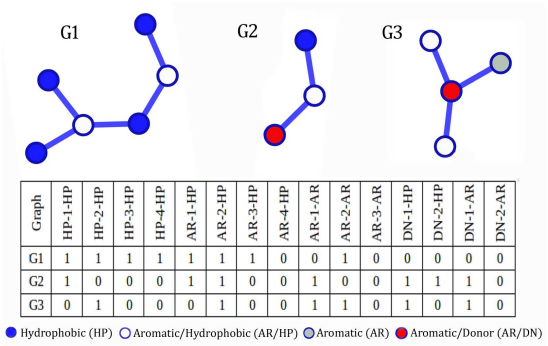


Fig. 2. PLI bipartite graph. Scheme of a graph depicting interactions at the interface of ricin protein and its ligand DDP (PDB id 1IL5 chain A). Protein nodes are colored in purple and ligand nodes in green.



Graph	HP-1-HP	HP-2-HP	HP-3-HP	HP-4-HP	AR-1-HP	AR-2-HP	AR-3-HP	AR-4-HP	AR-1-AR	AR-2-AR	AR-3-AR	DN-1-HP	DN-2-HP	DN-1-AR	DN-2-AR
G1	1	1	1	1	1	1	1	0	0	1	0	0	0	0	0
G2	1	0	0	0	1	1	0	0	1	0	0	1	1	1	0
G3	0	1	0	0	0	1	0	0	1	1	0	1	0	1	0

Fig. 3. Example of graphs and their feature vectors in the matrix.

well as the distance between them. Figure 3 provides an example of the proposed matrix as well as an example of a graph and how it is represented by a feature vector in this matrix. In the mentioned Figure, for instance, an attribute HP-4-HP means that the graph has two hydrophobic atoms separated by four edges of distance.

To perform dimensionality and noise reduction in the matrix that represents the PLI graphs, we used Singular Value Decompositon (SVD) (Demmel, 1997), which is a preprocessing step for the unsupervised learning task. SVD is widely used in data mining for this purpose.

We provide 3 options of clustering algorithms to the users, each based on a different paradigm. K-medoids (Kaufman and Rousseeuw, 1987) is centroid based, Agglomerative is hierarchical (Kaufman and Rousseeuw, 2009) and Spectral (Ng *et al.*, 2001) consists in algorithms that cluster data points using eigenvectors of matrices derived from the input data. Also, we provide 2 metrics to evaluate the quality of clustering: average silhouette width (asw) (Rousseeuw, 1987) and Calinski-Harabaz index (Caliński and Harabasz, 1974).When used with default parameters, visGReMLIN automatically selects K-medoids and average silhouette width. The output of *Unsupervised learning* block are clusters of similar PLI graphs.

2.2.3 Motif prediction

The *Motif prediction* block (Figure 1 - D) takes as input the clusters of PLI graphs and then a frequent subgraph mining (FSM) experiment is conducted to search for conserved motifs in each cluster.

The algorithm selected for FSM was gSpan (Yan and Han, 2002), which is a highly cited FSM algorithm. In accordance with Jiang *et al.* (2013), considering a graph dataset $D = \{G_0, G_1, ..., G_n\}$, $support(g)$ denotes the number of graphs in D which have g as a subgraph. The purpose of FSM is finding any subgraph g whose $support(g) \geq minSup$ (a minimum support threshold).

By default, visGReMLIN runs FSM with $support$ varying from 0.1 to 1.0 with step 0.1 (the amount in which the support varies) for each cluster. As support increases, we obtain subgraphs that are in a high fraction in the graph input dataset. Nonetheless, the number of total subgraphs tend to decrease. Additionally, as the $support$ increases, the resulting subgraphs tend to be small, which is expected as it is difficult to find big patterns in the whole graph input dataset.

Subgraphs obtained from FSM are filtered to keep only the maximal ones. Hence, in this paper, protein-ligand conserved motifs are maximal frequent subgraphs. Subgraphs resulting from FSM algorithms can be structurally repetitive, as a frequent subgraph can present other frequent subgraphs within it (Yan and Han, 2003). In accordance with Koyutürk *et al.* (2004), maximal frequent subgraphs are the most interesting ones in biological networks. Hence, visGReMLIN filters only the maximal subgraphs. There is no loss of information, as maximal subgraphs contains the discarded graphs.

The output of FSM algorithm reveals the frequent subgraphs and in which graphs from the input dataset they appear. However, the FSM does not provide a direct mapping from each node/edge of frequent subgraph to the corresponding node/edge in the graphs from input dataset. This mapping is interesting because allows users to exactly identify the patterns in the dataset analyzed. To conquer this shortness of FSM, we map maximal subgraphs to the input graph dataset through subgraph isomorphism

algorithm VF2 (Cordella *et al.*, 2004). The block *Motif prediction* has as output the PLI motifs and their mappings to the graph input dataset.

2.3 Data analytics and visualization

Data analytics and visualization block (Figure 1 - E) is comprised of 4 visualization modules that deliver the results of our strategy in a totally visual and interactive manner, allowing domain specialists to explore and make sense of conserved PLI motifs. visGReMLIN motifs can support users on gaining insights on which are the key factors responsible for molecular recognition in a specific dataset. Next, we detail the functionalities of each module.

2.3.1 Dataset details

In this module we present a table that summarizes the unsupervised learning results, as shown in Supplementary Material (Figure 2). The first column shows all groups and in the second column are listed all complexes of each group. Moreover, column one presents a *graph* icon that displays all ligands of a specific group and a *network* icon that shows all the input graphs for that group. Regarding the second column, each complex (PDB id and chain) is a link that directs the user to the structure on the PDB website. This module has also a text search that displays in the table only the lines that contain the entered characters.

2.3.2 Graph patterns table

A summary of PLI motifs is provided in *Graph patterns table*. As shown in the Supplementary Material (Figure 3), by selecting *Grouping columns*, we see a table in which the first column displays the *Motif size* (in number of nodes) segmented by the index of the group and the second column shows the occurrences of the respective motif segmented by support value. By occurrences we mean how many motifs were found with a specific size and considering a specific support. By clicking on the column label the table can be organized in ascending or descending order. Also, by clicking on group or support labels (colored lines inside the table), data can be organized in ascending or descending order. A set of filters are offered in the panel *Options* to explore this table. One can select a certain group to analyze using *Filter by group*, as well as to choose a minimum motif size with *Filter by minimum motif size* or a minimum number of motif occurrences with *Filter by minimum occurrences*.

By selecting *Simple table*, we see a heatmap table in which color is a pre-attentive attribute that encodes the frequency of motifs, as provided in the Supplementary Material (Figure 4). The darker the shade of blue, the higher the frequency. This table depicts the frequency and size (in number of nodes) of resulting motifs for each group (from unsupervised learning) and for each support value (from FSM). As the choice of support is empirical, we provide domain specialists with a panorama of the size, frequency, group and support to help them on deciding which is the appropriate support value to generate relevant and interesting motifs. There is a compromise between large and frequent motifs. The larger the motif, the smaller the frequency.

2.3.3 Motif view

Domain specialists can visually explore and make sense of PLI motifs through analytical interaction and navigation in *Motif view* module, presented in Figure 4. Motifs can be analyzed alone or in the context of the protein structure, which means that, given a motif, visGReMLIN highlights the motif in the graphs in which the motif was detected. Additionally, the tool presents motifs in schematic 2D representations as graphs or in the context of protein structure in a 3D molecular viewer. This module has 4 main panels that we detail next.

Options: provides a variety of filters to interact with the motifs. The common usage is selecting a support value and explore motifs using the filters below:

- View: displays PLI graphs without segmenting them by choosing *Free view*, or shows the PLI graphs in boxes containing PDB id, graph index, ligand name and group index by selecting *Pattern viewFree view*.
- Color nodes by: nodes can be colored according to atom type or molecule to which they belong (blue for protein and red for ligands).
- Filter by atom type: atoms of the selected type are highlighted (possible types are acceptor, aromatic, donor, hydrophobic, negative and positive).
- Filter by interaction type: interactions of the selected type are highlighted (possible types are aromatic stacking, hydrogen bond, hydrophobic, repulsive and salt bridge).
- Filter by group: displays only graphs of the selected groups from unsupervised learning.
- Filter by vertex number: shows those graphs that contains the selected number of nodes.
- Remove pattern selection: in case a motif has been selected in the panel *Motif graphs*, this option removes the selection.
- Search for a residue, ligand, or atom: highlights nodes from graphs that contain residue/ligand/atom in the text search.
- Show node labels: enables or disables the display of labels for the nodes.

Graphs legend: this panel presents the legends. Each atom type is associated with a specific color. The same holds for each interaction type.

Motif graphs: in this panel users can navigate through motifs, which are the frequent subgraphs from a dataset of protein-ligand complexes represented as graphs. By clicking on a motif, only PLI graphs that contain such motif are displayed on the panel *Input PLI graphs*. Additionally, the motif is highlighted in the context of the PLI graph. Figure 7-A provides an example of pattern selection. Types of atoms and interactions are displayed on demand by hovering the mouse over nodes and edges respectively. Just the motifs from groups selected in *Filter by group* are shown.

Input PLI graphs: PLI graphs depicting protein-ligand interface for a set of complexes are shown in this panel in accordance with filters from panel *Options* and with the motif chosen in *Motif graphs*. visGReMLIN shows, for each graph, PDB id and chain, graph index, ligand name and group index. By hovering the mouse over the graph, we see some details on demand, depending on the part of the graph:

- Protein nodes: residue name and number, atom name, chain, atom type. An example is provided in Figure 5-A
- Ligand nodes: ligand name and number, atom name, chain, atom type as shown in Figure 5-B.
- Edges: information about connected atoms, interaction type and distance between such atoms in angstroms (Å). Figure 5-C.

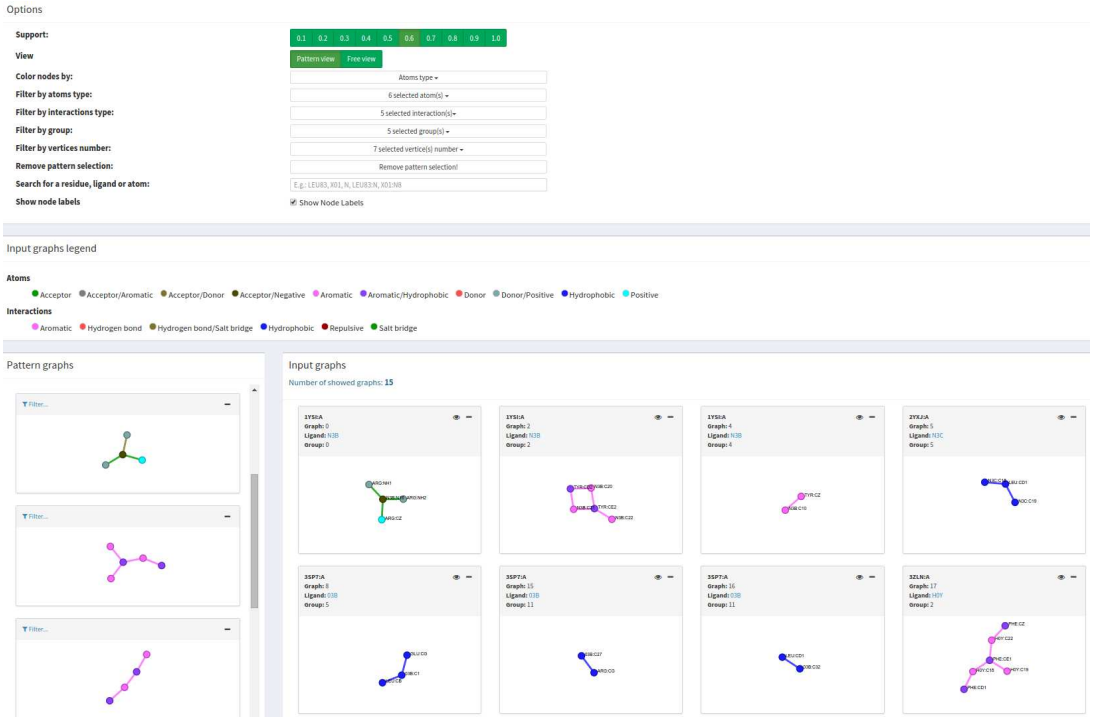


Fig. 4. Motif view. This module has 4 main panels: Options provides a set of filters to explore PLI motifs; Graphs legend displays colors used to depict atoms and edges; Motif graphs shows the set of PLI motifs according to selected filters; Input PLI graphs depicts PLI graphs that represent the interface between a ligand and a protein.

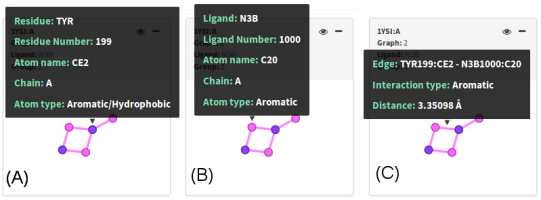


Fig. 5. Input PLI graphs. By hovering the mouse over nodes and edges we see some details on demand. In (A) we see details of a protein atom and in (B) we see details of a ligand atom. Edge details are shown in (C).

To support users on understanding and interpreting the patterns in the context of protein structure, we provide a 3D representation (Figure 7-C) of the protein-ligand interaction graphs in a molecule viewer (which we named *Interaction viewer*) by clicking on the *eye* icon. Also, a general 2D visualization for ligands (Figure 7-B) is provided by clicking on the ligand name in any graph from the subsection *Input PLI graphs*. Only ligands from graphs displayed in subsection *Input PLI graphs* are shown in the set of ligands and the ligand from the graph that the user clicked is highlighted in green.

2.3.4 Graphical analysis

visGReMLIN delivers an interactive interface to show a quantitative summary of motifs extracted from PLI graphs, as shown in Figure 5 from Supplementary Material. The common workflow in this panel is selecting the tab *Atoms type* or *Interactions type* and then choosing a physicochemical type of atoms or interactions to be displayed as histograms. In addition, histogram bars can be organized by *Support* value used in FSM or by *Group* from the unsupervised learning.

3 Results and discussion

Here we illustrate how visGReMLIN can be used to support domain specialists on gaining insights on which are the key factors for the interaction between a protein and a ligand.

Regarding the input parameters, we used default values for distance criteria, clustering algorithm (K-medoids) and evaluation metric (sillhouette coefficient). The support value selected was 0.6, which means that each motif was detected in at least 60% of PLI input graphs of each group. To show the generality and real-world applicability of our strategy, we used 2 datasets of protein-ligand complexes with different aims.

1. CDK: adapted from Schonbrunn *et al.* (2013), it consists in 73 protein-inhibitor complexes of the identical protein with varied ligands. This dataset is used to illustrate that our strategy can be used in a scenario involving protein promiscuity.
2. He: comprises 50 complexes involving ATP-binding proteins (complexed with ATP and ADP), thus having exact the similar ligands (ATP and ADP) complexed with varied proteins. This dataset was proposed by He *et al.* (2016) and it helps us to illustrate how our strategy can be used in a scenario involving ligand promiscuity.

The PDB ids of datasets used in both use cases are provided in Supplementary Material (Table 2).

3.1 CDK use case

Using CDK dataset as input, we created a project named CDK (available for access on visGReMLIN website), which resulted in 276 PLI input graphs divided in 11 groups, as shown in Supplementary Material (Figure 6). Here we present a qualitative analysis in which visGReMLIN motifs are compared to experimentally determined binding site residues/atoms of CDK interacting with the 2 most potent sulfonamide analogue inhibitors obtained in Schonbrunn *et al.* (2013). Residues/atoms determined as relevant in PLI in the mentioned work do not represent interactions between CDK and all its possible ligands in our dataset of CDK complexes. Thus it is expected that motifs found by visGReMLIN will not contain residues/atoms identical to those experimentally determined. Even so, we believe that the comparison is interesting to show that our strategy is able to detect the majority of relevant residues/atoms determined in Schonbrunn *et al.* (2013).

In Table 2, we show 26 relevant residues/atoms from CDK binding site. We searched each of these residues/atoms in visGReMLIN motifs and our strategy was able to find 18, which represents about 69%. Residue GLN85 was not considered here as it is involved in a water mediated interaction with ligand and our strategy does not consider this type of interaction. Residues in bold in Table 2 (LEU83; PHE 82; GLU81) are well known relevant residues for CDK as they are in the hinge region of this protein. It is important to point out that our strategy was able to detect all residue/atoms from CDK hinge region in motifs. In Figure 6 we show an example of atoms from hinge region detected in a visGReMLIN motif (PHE82:CE2 and PHE82:CZ).

Table 2. CDK binding site. Residues from CDK that interact with the 2 most potent sulfonamide analogue inhibitors.

Residue	Atom	visGReMLIN
ASP145	CB	✓
	CG	×
	OD1	✓
LYS33	CB	✓
	CD	✓
	CE	×
	CG	✓
	NZ	•
ASP86	N	✓
	CB	✓
	OD1	•
	OD2	✓
LYS89	CB	✓
	CE	×
	NZ	•
HIS84	O	×
LEU83	N	✓
	O	✓
PHE82	CE2	✓
	CZ	✓
GLU81	O	✓
PHE80	CB	✓
	CG	✓
	CD2	✓
	CE2	✓
	CZ	•

✓ Residues/atoms found in patterns;
• Found but not in patterns;
× Not found

Next we present examples of two interesting motifs detected by visually inspecting visGReMLIN results.

Figure 7-A depicts a motif (highlighted in green on the left hand side) with 5 nodes, 4 of them (in blue) are hydrophobic atoms and the other 1 (in purple) are aromatic/hydrofobic. This motif was found in 3 different complexes (3QRT.A, 3R8L.A, 3R9H.A) with different ligands (X14, Z30, Z67) of group 9 (which has 3 PLI graphs). These complexes are shown in Figure 7-A on the right hand side. Figure 7-B displays the 3 different ligands from PLI input graphs, with the ligand Z67 from complex 3R9H.A highlighted. Figure 7-C presents the motif in the context of the protein structure in a 3D molecular viewer for the complex 3R9H.A.

3.2 ATP use case

We created a project called ATP_50 (available for access on visGReMLIN website) in which we use data from He *et al.* (2016) as input. It resulted in 432 PLI graphs segmented in 5 groups, as show in Supplementary Material (Figure 7). Here we present a qualitative analysis in which we discuss some interesting motifs detected by visually inspecting GReMLIN results. Also, it is important to mention that from the 7 residues that He *et al.* (2016) considers as the motif for ATP dataset (VAL, GLY, THR, LEU, ALA, TYR, LYS) visGReMLIN was able to detect 3 (GLY, VAL, LYS) using default parameters. It is expected that visGReMLIN motifs differ from the mentioned work as they are computed in a totally different manner and with different purposes. While He *et al.* (2016) is focused on finding a unique motif (composed of residues) that summarizes the binding for a specific dataset (no matter the size of the dataset), our strategy aims at delivering frequent substructures (common arrangements of atoms) in PLI interaction. In visGReMLIN, motifs detected are coupled with a interactive visual tool to support on the understanding of such motifs, posing and answering questions about them.

We show an example of an interesting motif detected by visually inspecting visGReMLIN results.

Figure 6-A shows a 4-node motif (highlighted in green on the left hand side), 3 nodes (in brown) are acceptor/negative and the other node (in gray) is donor/positive. Such motif was found in 12 different complexes with 2 different ligands (ATP and ADP) of group 4 (which has 17 PLI graphs). One of these complexes (4AI6.A) are shown in Figure 6-A on the right hand side. Figure 6-B displays the ATP ligand. Figure 6-C presents the motif in the context of the protein structure in a 3D molecular viewer for the complex 4AI6.A.

4 Conclusion

In this paper we present visGReMLIN, a user-friendly web-server that brings together a computational strategy to detect motifs at protein-ligand interface and a visual interactive platform to explore and interpret such

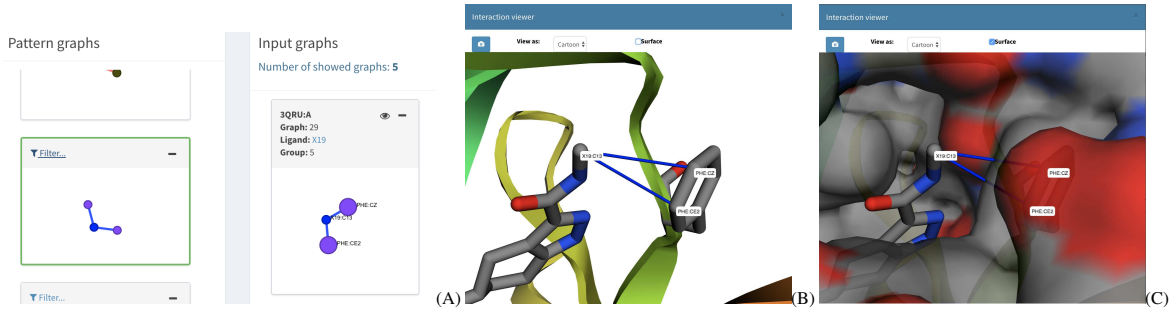


Fig. 6. Atoms from hinge region (PHE82:CE2 and PHE82:CZ) in a visGReMLIN motif. (A) depicts the motif (on the left hand side) and one of the PLI input graphs that contains the motif (on the right hand side). (B) shows the motif in the context of CDK structure and (C) displays exact the same structure in the same position of (B) but now visualized as surface in Interaction viewer.

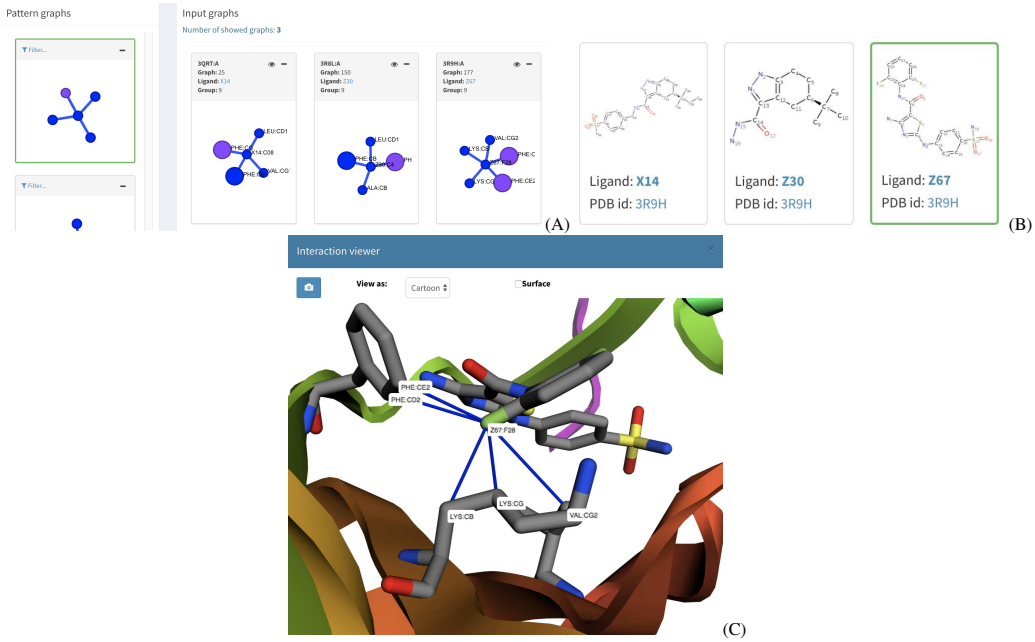


Fig. 7. 5-node motif from CDK dataset. The motif is displayed in (A), highlighted in green on the left hand side. PLI input graphs from group 9 in which the motif was detected are shown in the right hand side of (A). (B) provides images of ligands from PLI input graphs in group 9 (X14, Z30, Z67) with the ligand Z67 from complex 3R9H.A highlighted in green. (C) shows the motif in the complex 3R9H.A, in a 3D molecular viewer.

patterns. By motifs we mean frequent subgraphs detected at the interfaces between proteins and ligands. visGReMLIN motifs can support users on gaining insights on which are the key atoms/residues responsible for protein-ligand interaction in a dataset of complexes.

To illustrate the ability of our strategy on supporting users on the detection and understanding of motifs, we conducted 2 use cases. In the first one we used a dataset of 73 identical CDKs in complex with a varied set of ligands. We compared our motifs with experimental results to show that visGReMLIN is able to find relevant atoms/residues experimentally determined. In the second use case we used a dataset of 50 complexes that involve the ATP ligand in complex with different proteins. We compared our motifs with those found on a computational study to show that our strategy was able to detect the same relevant residues. Also, these use cases show that visGReMLIN can be used with a dataset of same/similar proteins in complex with different ligands and with a dataset of same/similar ligands in complex with different proteins, which means that the tool can be used in a scenario involving protein promiscuity and ligand promiscuity. We believe it is an important result of or work, as many available methods are limited to one of these scenarios.

As future work, we would like to investigate if motifs detected by visGReMLIN can be used to predict protein-ligand interaction. Considering that our strategy is able to characterize the interface motifs between a dataset of proteins and ligands, we are interested in using these motifs to help us to choose potential ligands for a specific protein. Also, we plan to systematically measure user insights and impressions about the motif detection and the proposed visualization to help us to improve visGReMLIN.

Funding

This work has been supported by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG).

References

Berman, H. M. *et al.* (2000). The protein data bank. *Nucleic acids research*, **28**(1), 235–242.

Bonham-Carter, O., Steele, J., and Bastola, D. (2013). Alignment-free genetic sequence comparisons: a review of recent approaches by word analysis. *Briefings in bioinformatics*, **15**(6), 890–905.

Caliński, T. and Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, **3**(1), 1–27.

Chandel, T. I., Zaman, M., Khan, M. V., Ali, M., Rabbani, G., Ishtikhar, M., and Khan, R. H. (2018). A mechanistic insight into protein-ligand interaction, folding, misfolding, aggregation and inhibition of protein aggregates: An overview. *International Journal of Biological Macromolecules*, **106**, 1115 – 1129.

Cobanoglu, M. C., Liu, C., Hu, F., Oltvai, Z. N., and Bahar, I. (2013). Predicting drug–target interactions using probabilistic matrix factorization. *Journal of chemical information and modeling*, **53**(12), 3399–3409.

Cordella, L. P., Foggia, P., Sansone, C., and Vento, M. (2004). A (sub) graph isomorphism algorithm for matching large graphs. *IEEE transactions on pattern analysis and machine intelligence*, **26**(10), 1367–1372.

da Silveira, C. H., Pires, D. E., Minardi, R. C., Ribeiro, C., Veloso, C. J., Lopes, J. C., Meira, W., Neshich, G., Ramos, C. H., Habesch, R., *et al.* (2009). Protein cutoff scanning: A comparative analysis of cutoff dependent and cutoff free methods for prospecting contacts in proteins. *Proteins: Structure, Function, and Bioinformatics*, **74**(3), 727–743.

Demmel (1997). Applied numerical linear algebra.

Desaphy, J., Raimbaud, E., Ducrot, P., and Rognan, D. (2013). Encoding protein–ligand interaction patterns in fingerprints and graphs. *Journal of chemical information and modeling*, **53**(3), 623–637.

Diestel, R. (2000). *Graph theory*. © Springer-Verlag New York.

Fassio, A. V., Martins, P. M., Guimarães, S. d. S., Junior, S. S., Ribeiro, V. S., de Melo-Minardi, R. C., and Silveira, S. d. A. (2017). Vermont: a multi-perspective visual interactive platform for mutational analysis. *BMC bioinformatics*, **18**(10), 403.

Fassio, A. V., Santos, L. H. S., Silveira, S. A., Ferreira, R. S., and de Melo-Minardi, R. C. (2018). nAPOLI: a graph-based strategy to detect and visualize conserved protein-ligand interactions in large-scale. To be published.

Gao, M. and Skolnick, J. (2013a). Apoc: large-scale identification of similar protein pockets. *Bioinformatics*, **29**(5), 597–604.

Gao, M. and Skolnick, J. (2013b). A comprehensive survey of small-molecule binding pockets in proteins. *PLoS computational biology*, **9**(10), e1003302.

Gonçalves-Almeida, V. M., Pires, D. E., de Melo-Minardi, R. C., da Silveira, C. H., Meira, W., and Santoro, M. M. (2011). Hydropace: understanding and predicting cross-inhibition in serine proteases through hydrophobic patch centroids. *Bioinformatics*, **28**(3), 342–349.

He, W., Liang, Z., Teng, M., and Niu, L. (2016). Libme: automatic extraction of 3d protein–ligand binding motifs for mechanistic analysis of protein–ligand recognition. *FEBS open bio*.

Jiang, C., Coenen, F., and Zito, M. (2013). A survey of frequent subgraph mining algorithms. *The Knowledge Engineering Review*, **28**(1), 75–105.

Kadukova, M. and Grudinin, S. (2017). Convex-pl: a novel knowledge-based potential for protein-ligand interactions deduced from structural databases using convex optimization. *Journal of Computer-Aided Molecular Design*, **31**(10), 943–958.

Kaufman, L. and Rousseeuw, P. (1987). *Clustering by means of medoids*. North-Holland.

Kaufman, L. and Rousseeuw, P. J. (2009). *Finding groups in data: an introduction to cluster analysis*, volume 344. John Wiley & Sons.

Koyutürk, M. *et al.* (2004). An efficient algorithm for detecting frequent subgraphs in biological networks. *Bioinformatics*, **20**(suppl 1), i200–i207.

Kufareva, I., Ilatovskiy, A. V., and Abagyan, R. (2011). Pocketome: an encyclopedia of small-molecule binding sites in 4d. *Nucleic acids research*, **40**(D1), D535–D540.

Kuttner, Y. Y., Sobolev, V., Raskind, A., and Edelman, M. (2003). A consensus-binding structure for adenine at the atomic level permits searching for the ligand site in a wide spectrum of adenine-containing complexes. *Proteins: Structure, Function, and Bioinformatics*, **52**(3), 400–411.

Nakadai, M., Tomida, S., and Sekimizu, K. (2016). An intriguing correlation based on the superimposition of residue pairs with inhibitors that target protein-protein interfaces. *Scientific reports*, **6**.

Nebel, J.-C., Herzyk, P., and Gilbert, D. R. (2007). Automatic generation of 3d motifs for classification of protein binding sites. *BMC bioinformatics*, **8**(1), 321.

Ng, A. Y., Jordan, M. I., and Weiss, Y. (2001). On spectral clustering: Analysis and an algorithm. In *ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS*, pages 849–856. MIT Press.

Nobeli, I., Favia, A. D., and Thornton, J. M. (2009). Protein promiscuity and its implications for biotechnology. *Nature biotechnology*, **27**(2), 157–167.

Pai, P. P., Dattatreya, R. K., and Mondal, S. (2017). Ensemble Architecture for Prediction of Enzyme-ligand Binding Residues Using Evolutionary Information. *MOLECULAR INFORMATICS*, **36**(11).

Pires, D. E., de Melo-Minardi, R. C., dos Santos, M. A., da Silveira, C. H., Santoro, M. M., and Meira, W. (2011). Cutoff scanning matrix (csm): structural classification and function prediction by protein inter-residue distance patterns. *BMC genomics*, **12**(4), S12.

Pires, D. E., de Melo-Minardi, R. C., da Silveira, C. H., Campos, F. F., and Meira Jr, W. (2013). acsm: noise-free graph-based signatures to large-scale receptor-based ligand prediction. *Bioinformatics*, **29**(7), 855–861.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, **20**, 53–65.

Santana, C. A., Cerqueira, F. R., da Silveira, C. H., Fassio, A. V., de Melo-Minardi, R. C., and Silveira, S. d. A. (2016). Gremlin: A graph mining strategy to infer protein-ligand interaction patterns. In *Bioinformatics and Bioengineering (BIBE), 2016 IEEE 16th International Conference on*, pages 28–35. IEEE.

Schonbrunn, E., Betzi, S., Alam, R., Martin, M. P., Becker, A., Han, H., Francis, R., Chakrasali, R., Jakkaraj, S., Kazi, A., *et al.* (2013). Development of highly potent and selective diaminothiazole inhibitors of cyclin-dependent kinases. *Journal of medicinal chemistry*, **56**(10), 3768–3782.

Silveira, S. A., Fassio, A. V., Gonçalves-Almeida, V. M., de Lima, E. B., Barcelos, Y. T., Aburjaile, F. F., Rodrigues, L. M., Meira Jr, W., and de Melo-Minardi, R. C. (2014). Vermont: Visualizing mutations and their effects on protein physicochemical and topological property conservation. In *BMC proceedings*, volume 8, page S4. BioMed Central.

Tuncbag, N., Gursoy, A., and Keskin, O. (2011). Prediction of protein–protein interactions: unifying evolution and structure at protein interfaces. *Physical biology*, **8**(3), 035006.

Vinga, S. (2014). Alignment-free methods in computational biology. *Briefings in bioinformatics*, **15**(3), 341–342.

Wang, S., Ma, J., Peng, J., and Xu, J. (2013). Protein structure alignment beyond spatial proximity. *Scientific reports*, **3**.

Woińska, M., Grabowsky, S., Dominiak, P. M., Woźniak, K., and Jayatilaka, D. (2016). Hydrogen atoms can be located accurately and precisely by x-ray crystallography. *Science advances*, **2**(5), e1600192.

Yan, X. and Han, J. (2002). gspan: Graph-based substructure pattern mining. In *Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on*, pages 721–724. IEEE.

Yan, X. and Han, J. (2003). Closegraph: mining closed frequent graph patterns. In *Proceedings of the ninth ACM SIGKDD*, pages 286–295. ACM.

Capítulo 4

Conclusões

O número cada vez maior de dados biológicos disponíveis, como informações sobre sequências de resíduos de uma proteína ou sua estrutura tridimensional, torna necessário o uso de ferramentas computacionais que possibilitem a análise e o estudo desta crescente quantidade de dados. Neste trabalho, apresentamos duas abordagens interativas visuais para investigar diferentes questões biológicas. Com o VERMONT, disponibilizamos uma estratégia de avaliação de impacto de mutações que combina o cálculo e a visualização de interações intramoleculares, de grau de acessibilidade a solvente, propriedades físico-químicas de resíduos e métricas de redes complexas. Para verificar sua aplicabilidade analisamos mutações experimentalmente já estudadas, verificando que para o caso estudado, mutações danosas tendem a ter valores baixos e conservados de acessibilidade a solvente com um número conservado de interações hidrofóbicas.

No visGReMLIN, buscamos padrões de interações não covalentes na interface entre proteínas e ligantes. Sua abordagem consiste em modelar as interações como grafos em nível atômico, agrupar grafos similares e realizar uma mineração de subgrafos frequentes em cada um dos grupos. A interface visual permite visualizar as interações em 2 ou 3 dimensões, obter informações sobre os grupos de grafos, os padrões encontrados e a correspondência destes padrões nos grafos de entrada. Diversos filtros também estão disponíveis para explorar informações de interesse do pesquisador. Nos casos avaliados neste trabalho, o visGReMLIN foi capaz de identificar com alto grau de sucesso, átomos e resíduos importantes já analisados

experimentalmente ou computacionalmente. Ambas as ferramentas podem fornecer informações úteis a pesquisadores interessados no estudo de proteínas, com a vantagem de permitir uma análise visual integrada dos resultados encontrados.

Referências Bibliográficas

CGAL (2017). The cgal project. cgal user and reference manual [internet]. 4.10.

Chandel, T. I. et al. (2017). A mechanistic insight into protein ligand interaction, folding, misfolding, aggregation and inhibition of protein aggregates: An overview. *International journal of biological macromolecules*, 31(10):943–958.

Chen, C.-W.; Lin, J. & Chu, Y.-W. (2013). istable: off-the-shelf predictor integration for predicting protein stability changes. *BMC bioinformatics*, p. S5.

Cordella, L. P. et al. (2004). A (sub) graph isomorphism algorithm for matching large graphs. *IEEE transactions on pattern analysis and machine intelligence*, 26(10):1367–1372.

Desaphy, J.; Raimbaud, E.; Ducrot, P. & Rognan, D. (2013). Encoding proteinâligand interaction patterns in fingerprints and graphs. journal of chemical information and modeling. *Journal of chemical information and modeling*, 53(3):623â637.

Dunn, M. F. (2007). Protein ligand interactions: general description. *eLS*.

Furnham, N. et al. (2014). The catalytic site atlas 2.0: cataloging catalytic sites and residues identified in enzymes. *Nucleic acids research*, 42(D1):D485–D489.

Gao, M. & Skolnick, J. (2013a). Apoc: large-scale identification of similar protein pockets. *Bioinformatics*, 29(5):597–604.

Gao, M. & Skolnick, J. (2013b). A comprehensive survey of small-molecule binding pockets in proteins. *PLoS computational biology*, 9(10):e1003302.

Giollo, M. et al. (2014). Neemo: a method using residue interaction networks to improve prediction of protein stability upon mutation. *BMC genomics*, 15(4):S7.

Gonçalves-Almeida, V. M. et al. (2011). Hydropace: understanding and predicting cross-inhibition in serine proteases through hydrophobic patch centroids. *Bioinformatics*, 28(3):342–349.

He, W.; Liang, Z.; Teng, M. & Niu, L. (2016). Libme: automatic extraction of 3d protein ligand binding motifs for mechanistic analysis of proteinâligand recognition. *FEBS open bio*, 6(12):1331–1340.

Kadukova, M. & Grudinin, S. (2017). Convex-pl: a novel knowledge-based potential for protein ligand interactions deduced from structural databases using convex optimization. *Journal of Computer-Aided Molecular Design*, 31(10):943–958.

Kuttner, Y. Y.; Sobolev, V.; Raskind, A. & Edelman, M. (2003). A consensus binding structure for adenine at the atomic level permits searching for the ligand site in a wide spectrum of adenine-containing complexes. *Proteins: Structure, Function, and Bioinformatics,*, 52(3):400–411.

Laimer, J. et al. (2015). Maestro-multi agent stability prediction upon point mutations. *BMC bioinformatics*, 16(1):116.

LEE, Byungkook; RICHARDS, F. M. (1971). The interpretation of protein structures: estimation of static accessibility. *Journal of molecular biology*, 55(3):379–IN4.

Luscombe et al. (2001). What is bioinformatics? an introduction and overview. *Department of Molecular Biophysics and Biochemistry Yale University New Haven, USA*.

Mancini, A. et al. (2004). Sting contacts: a web-based application for identification and analysis of amino acid contacts within protein structure and across protein interfaces. *Bioinformatics*, 20(13):2145–2147.

Nair, A. S. (2007). Computational biology and bioinformatics: A gentle overview. *Communications of the Computer Society of India*.

Nebel, J.; Herzyk, P. & Gilbert, D. R. (2007). Automatic generation of 3d motifs for classification of protein binding sites. bmc bioinformatics. *BMC bioinformatics*, 8(1):321.

Okabe, A. et al. (2008). *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams, Second Edition*. John Wiley & Sons Ltd.

Pai, P. P.; Dattatreya, R. K. & Mondal, S. (2017). Ensemble architecture for prediction of enzyme-ligand binding residues using evolutionary information. *Molecular Informatics*, 36(11).

Pires, D. E.; Ascher, D. B. & Blundell, T. L. (2014). Duet: a server for predicting effects of mutations on protein stability using an integrated computational approach. *Nucleic acids research*, 42(W1):W314–W319.

Pires, D. E.; de Melo-Minardi, R. C.; da Silveira, C. H.; Campos, F. F. & Meira Jr, W. (2013). acsm: noise-free graph-based signatures to large-scale receptor-based ligand prediction. *Bioinformatics*, 29(7):855–861.

Santana, C. A. et al. (2016). Gremlin: A graph mining strategy to infer protein ligand interaction patterns. Em *Bioinformatics and Bioengineering (BIBE), 2016 IEEE 16th International Conference on*, pp. 28–35. IEEE.

Shatsky, M. et al. (2014). A method for simultaneous alignment of multiple protein structures. *Proteins: Structure, Function, and Bioinformatics*, 56(1):S4.

SILVEIRA, S. A. et al. (2014). Vermont: Visualizing mutations and their effects on protein physicochemical and topological property conservation. *BMC proceedings*, p. S4.

Topham, C. M.; Srinivasan, N. & Blundell, T. L. (1997). Prediction of the stability of protein mutants based on structural environment-dependent amino acid substitution and propensity tables. *Protein Engineering*, 10(1):7–21.

Tuncbag, N.; Gursoy, A. & Keskim, O. (2011). Prediction of proteinâprotein interactions: unifying evolution and structure at protein interfaces. *Physical biology*, 8(3):035006.

Xing, D. et al. (2016). Insights into protein ligand interactions: mechanisms, models, and methods. *International journal of molecular sciences*, 17(2):144.

Apêndice A

Arquivo suplementar do artigo 1

ADDITIONAL FILE 1

Vermont: a multi-perspective visual interactive platform for mutational analysis

Alexandre V Fassio[1,2*†], Pedro M Martins[1,2†], Samuel da S Guimarães[3], Sócrates S A Junior[3], Vagner S Ribeiro[3], Raquel C de Melo-Minardi[1] and Sabrina de A Silveira[3]

In this document we present additional details and figures about VERMONT platform. The document is organized in sections that are correspondent to those in the main document.

Methods

Input module

Figure 1 shows VERMONT input module.

Topological properties module

Here we comment on some uses of graphs and network measures in a biological context. Bongo[1] uses graphs to represent residue-residue interaction networks and to assign key residues that are important for maintaining such networks. Also, they applied a graph theory concept, vertex cover, which identifies key residues for analyzing structural effects of single point mutations. In [2], complex networks were used to study the role of a residue in local and global structures. High betweenness is expected for key residues that act as a bridge in protein structure, such as those that bring together two different secondary structures. Closeness, in turn, could indicate the functional role of a residue. Also in [2], high closeness values were observed for disease-associated nsSNPs.

In VERMONT, three common complex network centrality measures were computed for each residue. Next, we describe them in detail.

- Degree: the degree of a vertex in a graph is the number of edges connected to it. For an undirected graph of n vertices, the degree k_i of a vertex i can be written in terms of the adjacency matrix as $k_i = \sum_{j=1}^{n} A_{ij}$.
- Betweenness: measures the extent to which a vertex lies on paths between other vertices. Let n_i to be the number of geodesic paths from vertex s to vertex t that pass through vertex i. Let g_{st} to be the total number of geodesic paths from s to t. Then the betweenness centrality of i is $x_i = \sum_{st} \frac{n_{st}^i}{g_{st}}$.
- Closeness: measures the mean distance from a vertex to all other vertices. Let d_i be the length of a geodesic path from i to j, meaning the number of edges along the path. Then the mean geodesic distance from vertex i to vertex j, averaged over all vertices j in the network, is $l_i = \frac{1}{n}\sum_j d_{ij}$. The mean l_i is not a centrality measure since it gives low values for central vertices and high values for less central ones. Therefore, the closeness centrality C_i is the inverse of l_i: $C_i = \frac{1}{l_i}$.

Energy variation prediction

The effects caused by a mutation can be evaluated through the calculation of Gibbs free energy change ($\Delta\Delta G$). Bearing this in mind, we combined energy variation with our visualization modules to potentialize the analysis of specialists. Currently, the prediction of the effect of each mutation is performed with the FoldX tool. To do so, the wild structure was defined as the input, and the FoldX default parameters for pH (7) and temperature (298 K) were used.

Departing from the standard deviation (0.46 kcal/mol) [3], we divided the effects into seven categories in which values range from highly stabilizing ($\Delta\Delta G <$ -1.84 kcal/mol), stabilizing (-1.84 kcal/mol $\leq \Delta\Delta G <$ -0.92 kcal/mol), slightly stabilizing (-0.92 kcal/mol $\leq \Delta\Delta G <$ -0.46 kcal/mol), neutral (-0.46 kcal/mol $< \Delta\Delta G \leq$ 0.46 kcal/mol), slightly destabilizing (+0.46 kcal/mol $< \Delta\Delta G \leq$ +0.92 kcal/mol), destabilizing (+0.92 kcal/mol $< \Delta\Delta G \leq$ +1.84 kcal/mol) and highly destabilizing ($\Delta\Delta G >$ 1.84 kcal/mol).

Results and Discussion

In this section, we use VERMONT to visually analyze 6 disease-associated mutations in tumor suppressor protein, p53, experimentally studied by Fersht and co-workers [4, 5]. We discuss a total of 8 mutations, showed in Table 1, 2 illustrative cases are in the main paper and the other 6 are in the *Additional file 1* due to space limitations. Also, some additional figures for

*Correspondence: alexandrefassio@dcc.ufmg.br
[1] Department of Computer Science, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil
[2] Department of Biochemistry and Immunology, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil
Full list of author information is available at the end of the article
†Equal contributor

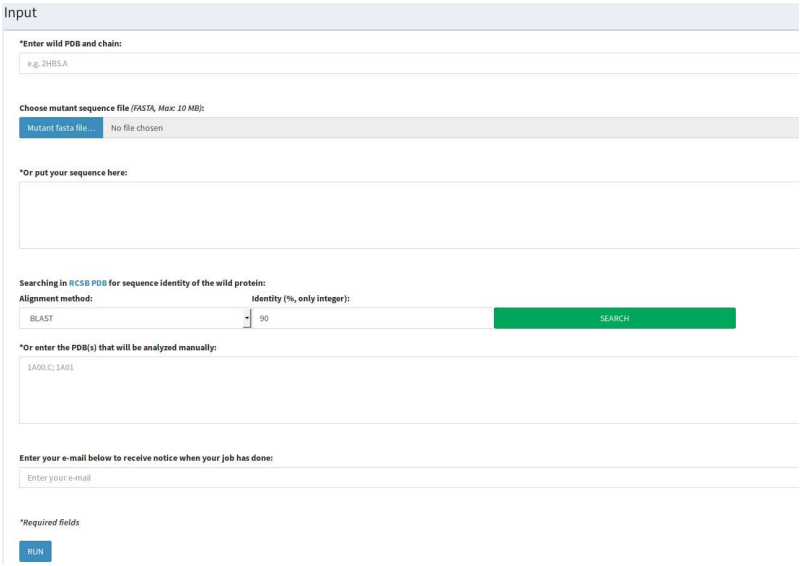


Figure 1 VERMONT *Input* module. It takes as input the PDB id and chain of a wild protein, the FASTA sequence of the respective mutant protein and a set of protein structures similar to the wild protein.

discussions that are in the main paper are showed in this section:

- Figure 2 shows the conservation on alignment position 180 (mutation Arg273His in protein p53) on the *Structure based sequence alignment module*.
- Figure 3 provides the topological properties for mutation Arg273His in protein p53. Alignment position 180, which corresponds to this mutation, is highlighted.
- Figure 4 shows interactions for Arg270 of 3EXL.A, which is in the alignment position 180 (related to Arg273His in protein p53). Interactions are displayed in a 3D molecule viewer and in a 2D graph.
- Figure 5 provides the topological properties for mutation Ile195Thr in protein p53. Alignment position 102, which corresponds to this mutation, is highlighted.

Use case

VERMONT input parameters were (i) PDB id 1TSR.A as wild protein; (ii) the mutant fasta file was generated by manually changing original residues in 1TSR.A fasta file by those that are the result of mutation; (iii) PSI-BLAST as alignment method; (iv) 70% of identity. The results are available to be explored and analyzed in VERMONT[1].

Complex network centrality measures for mutations Arg273His (no structural effects) and Ile195Thr (highly destabilising), discussed in the main paper, are presented in Figures 3 and 5, respectively.

Gly245Ser mutation corresponds to position 152 in the structural alignment, and it is non-conservative as Gly is nonpolar aliphatic and Ser is polar neutral. This column is highly conserved in the structural alignment, as Gly is present at 94% of the proteins. The accessibility is conserved and has low values (3 up to 48.6), as the whole column presents the same shade of gray. With regard to the topological properties, the degree is conserved (2 to 5); the betweenness is low (light shades in the column) and not well conserved, as the color is not very similar in the whole column; closeness is relatively conserved. Inspecting the interactions established in the alignment position 152, we see there are only hydrogen bonds, except in the PDB 2BIO.A that also presents a hydrophobic interaction. Considering all these aspects, we tend to point this mutation as probably damaging as it is non-conservative and has low and conserved values for accessibility, which means residues in this position are not exposed to solvent, being in the protein core, where we believe a mutation tends to have more impact on protein stability. This conclusion is in accordance with FoldX, which outlines this position with a red rectangle.

Arg249Ser, which is represented at position 156 in the structural alignment, is a non-conservative mutation as Arg is polar positive and Ser is polar neutral. The position 156 is highly conserved in the structural alignment as 90% of the residues are Arg. The accessibility is conserved and relatively low (9.5 up to 41.9)

[1] http://bioinfo.dcc.ufmg.br/vermont/results/view/case_study1/alignment

with a shade of gray in the whole column. Inspecting the topological properties, the degree and betweenness are not well conserved, as the column does not present a homogeneous shade; closeness is relatively conserved. Regarding the interactions, about 90% of the residues establish charged interactions, of which all are Arginines. Hydrogen bonds are highly conserved, being established by 99% of the residues, while hydrophobic interactions are relatively conserved in this column as 64% of the residues established this interaction type. Although Serine is also able to establish hydrogen bonds, the high conservation of charged interactions in this column indicates that Arginine likely further stabilize the protein. Thus, we would point this mutation as likely damaging because it is non-conservative, with low and conserved accessibility, despite FoldX points out this mutation as neutral.

Arg248Ala, which is represented in the structural alignment position 155, is a non-conservative mutation as Arg is polar positive and Ala is nonpolar aliphatic. This column is highly conserved in the structural alignment, presenting only Arginines. The accessibility is relatively high (27.3 up to 89.4) and conserved, with the whole column in a light shade of blue. With regard to the topological properties, the degree is well conserved (values 2 and 4); betweenness is not conserved; closeness is relatively conserved. When it comes to the interactions in position 155, all residues establish hydrogen bonds, so this interaction is highly conserved. Charged attractive interactions are not conserved as only 1 residue establishes this type of interaction. It is noteworthy that this mutation occurs in the DNA binding site (Figure 6), therefore the Arg248Ala mutation would likely diminish the protein-DNA affinity. Therefore, we consider this mutation as probably damaging due to its position, what is also confirmed by the high frequency of Arginines in this column. Bearing this in mind, we believe FoldX pointed out such mutation as neutral because it did not take the binding site into consideration.

Cys242Ser mutation corresponds to structural alignment position 149, and it is non-conservative as Cys is a residue with special properties (it can establish disulfide bridge) and Ser is polar neutral. The position 149 is highly conserved in the structural alignment with Cysteine residues. There is only one row, PDB id 2P52.A, that presents Ser (S). The accessibility is conserved and present low values (7.6 up to 38.6) having a light shade of gray, the only exception being 2P52.A (accessibility 62.7), which we consider as an outlier. Considering the topological properties, degree is well conserved (2 up to 5); betweenness is not conserved; closeness is relatively conserved. Regarding the interactions of alignment position 149, all residues establish hydrogen bonds, which are highly conserved, and 32 residues establish hydrophobic interaction. Having these aspects in mind, we consider this mutation as damaging as it is non-conservative (changing a cysteine, which is a residue with special properties) and it occurs in a position with low and conserved accessibility. On the other hand, FoldX points this mutation as slightly stabilizing. We further investigated Cys242 and discovered that it helps to stabilize p53 through a coordination system together with Zinc, Histidine and two other Cysteines [4, 5]. Therefore, Cys242Ser mutation is indeed destabilizing.

His168Arg mutation is represented in the structural alignment position 75, and it is conservative as both residues are polar positive. The alignment position 75 seems highly conserved, as 93% of the residues are Histidines, and the remaining residues are Arginines. The accessibility is relatively low and conserved (5.1 up to 39.8). Regarding the topological properties, the degree is relatively conserved (3 up to 7); betweenness is not well conserved; closeness is relatively conserved, being in a region with a light shade of yellow. When it comes to the interactions of position 75, all residues establish hydrogen bond interactions, while 82%, 85% and 87% of the residues establish charged attractive, charged repulsive and hydrophobic interactions, respectively, which are well conserved. Although FoldX points out this mutation as neutral, we consider this mutation as likely damaging because accessibility is relatively low and conserved, and the interactions are well conserved. As showed in [5], the Histidine substitution produced a distortion around the mutation site, what caused the residues 166-170 to be omitted in the solved structure (PDB 2BIN) (Figure 7). The authors also showed that the combination of both His168Arg and Arg249Ser mutation reversed the structural changes induced by these single mutations. In fact, all Arginines we observed in the position 75 appeared only when the Arg249Ser mutation occurred (position 156), what further confirms that the single His168Arg mutation is damaging.

Val143Ala, which is in the position 50 in the structural alignment, is conservative as Val and Ala are both nonpolar aliphatic. Column 50 is highly conserved, presenting only Valines, except 1 row (2J1W.A) that presents an Alanine. The accessibility is very low and conserved (0 up to 4.3). Considering the topological properties, degree is relatively conserved (3 up to 5); betweenness and closeness are relatively conserved. The hydrogen bonds and hydrophobic interactions in position 50 are highly conserved, as 100% and 91% of the residues, respectively, establish these interactions. Considering all these aspects, we tend to point out Val143Ala as damaging, because the position 50 presents very low and conserved accessibility

with highly conserved hydrophobic interactions, being a mutation in the protein core, which we believe have an impact on stability. Moreover, according to Lesk color scheme, the mutation is non-conservative, as Val is hydrophobic and Ala is small nonpolar. In fact, Val143Ala is a mutation which results in a residue with smaller volume (Ala). Our conclusion is in accordance with FoldX, which outlines this position with a red rectangle.

Availability of data and material
Vermont interactive platform and *Additional file 1* are available at: http://bioinfo.dcc.ufmg.br/vermont/

Competing interests
The authors declare that they have no competing interests.

Funding
This work has been supported by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG).

Author's contributions
SAS and RCM conceived the VERMONT platform. AVF and PMM designed and implemented the tool. SSG, SSA, and VSR implemented algorithms for property computation. SAS and RCM analyzed the results and wrote the manuscript. All authors read and approved the final manuscript.

Author details
[1] Department of Computer Science, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil. [2] Department of Biochemistry and Immunology, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil. [3] Computer Science Department, Universidade Federal de Viçosa, Viçosa, Brazil.

References
1. Cheng, T.M., Lu, Y.-E., Vendruscolo, M., Blundell, T.L., *et al.*: Prediction by graph theoretic measures of structural effects in proteins arising from non-synonymous single nucleotide polymorphisms. PLoS Comput Biol **4**(7), 1000135 (2008)
2. Li, Y., Wen, Z., Xiao, J., Yin, H., Yu, L., Yang, L., Li, M.: Predicting disease-associated substitution of a single amino acid by analyzing residue interactions. BMC bioinformatics **12**(1), 14 (2011)
3. Energy range. https://evosite3d.blogspot.com.br/2015/03/tutorial-estimating-stability-effect-of.html?m=1
4. Friedler, A., Veprintsev, D.B., Hansson, L.O., Fersht, A.R.: Kinetic instability of p53 core domain mutants implications for rescue by small molecules. Journal of Biological Chemistry **278**(26), 24108–24112 (2003)
5. Joerger, A.C., Ang, H.C., Veprintsev, D.B., Blair, C.M., Fersht, A.R.: Structures of p53 cancer mutants and mechanism of rescue by second-site suppressor mutations. Journal of Biological Chemistry **280**(16), 16030–16037 (2005)
6. Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., Ferrin, T.E.: UCSF Chimera–a visualization system for exploratory research and analysis. J Comput Chem **25**(13), 1605–1612 (2004)

Figure 2 *Residue conservation highlighted for alignment position 180, which corresponds to the conservative mutation Arg273His in protein p53 (1TSR.A). This position is highly conserved using the CINEMA color scheme, with about 89% of Arg and 5% of His and Cys each.*

Table 1 Mutations (nsSNPs) in the p53 (PDBid 1TSR) core domain that were experimentally characterized.

Mutant categories	Mutations
No structural effects	Arg273His
Weakly/locally destabilising	Gly245Ser Arg249Ser Arg248Ala
Highly destabilising/global unfolding	Cys242Ser His168Arg Val143Ala Ile195Thr

Figure 3 *Topological properties (network centrality measures) highlighted for alignment position 180, which corresponds to mutation Arg273His in protein p53 (1TSR.A). (A) Degree heatmap. Degree is conserved as the highlighted column presents a similar shade of orange. (B) Betweenness heatmap. Column 180 presents many different shades of blue, which means that betweenness is not conserved. (C) Closeness heatmap. This measure is relatively conserved as column 180 presents quite similar shades of yellow. Note that closeness has conserved regions.*

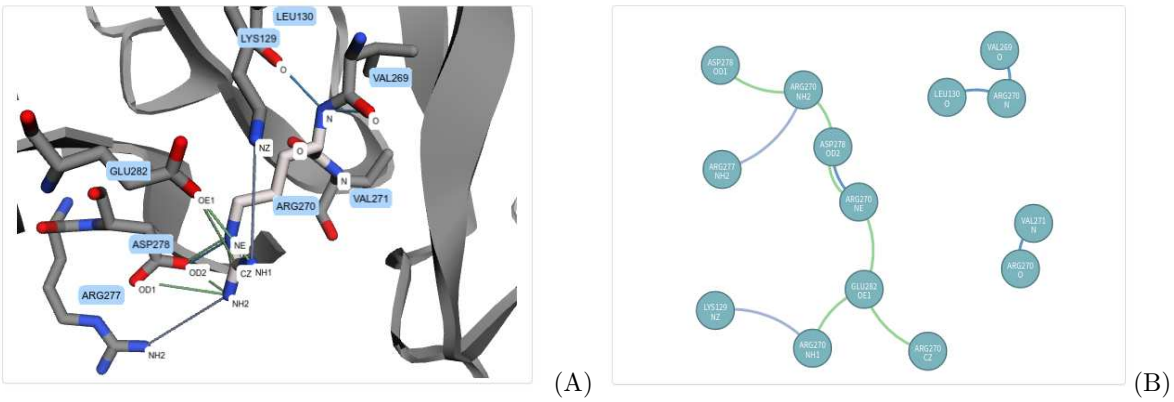


Figure 4 Interactions established by residue Arg270 of 3EXL.A. This residue is in the alignment position 180, which corresponds to mutation Arg273His in protein p53 (1TSR.A). By clicking on any residue on the *Interactions module*, VERMONT shows its interactions in a 3D molecule viewer and in a 2D graph. Here we illustrate how the domain specialist can inspect the interactions established by each particular residue in a detailed manner in the atomic level. (A) 3D interactions for Arg270 of 3EXL.A in a molecule viewer. (B) 2D graph schematic representation for interactions from Arg270 of 3EXL.A.

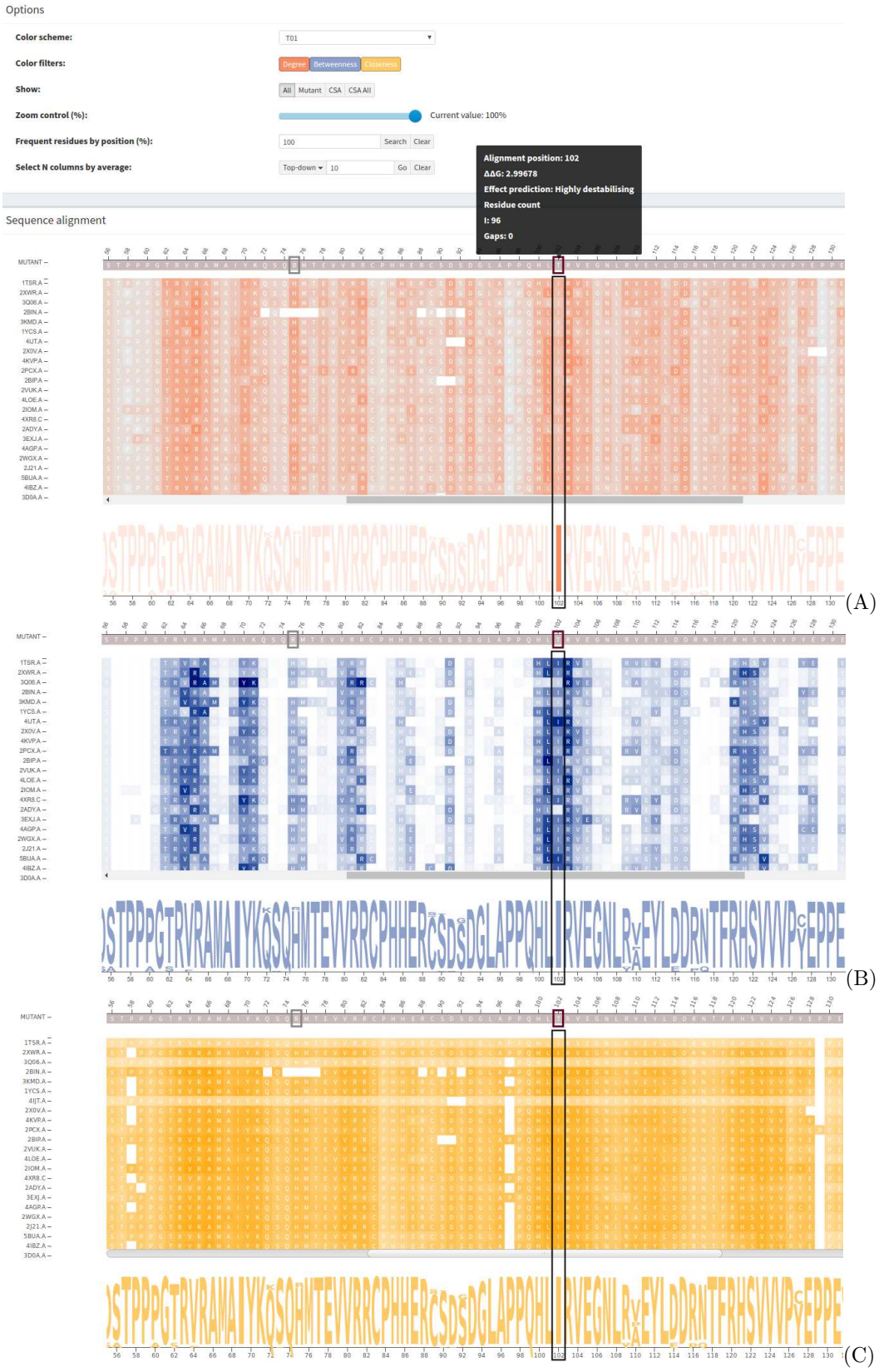
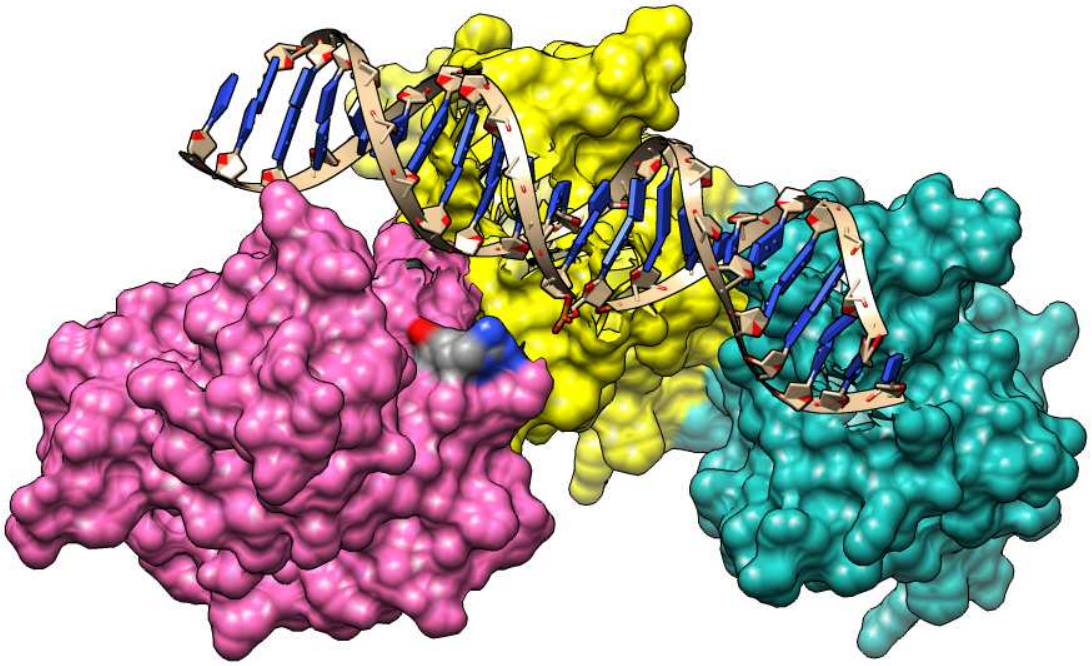
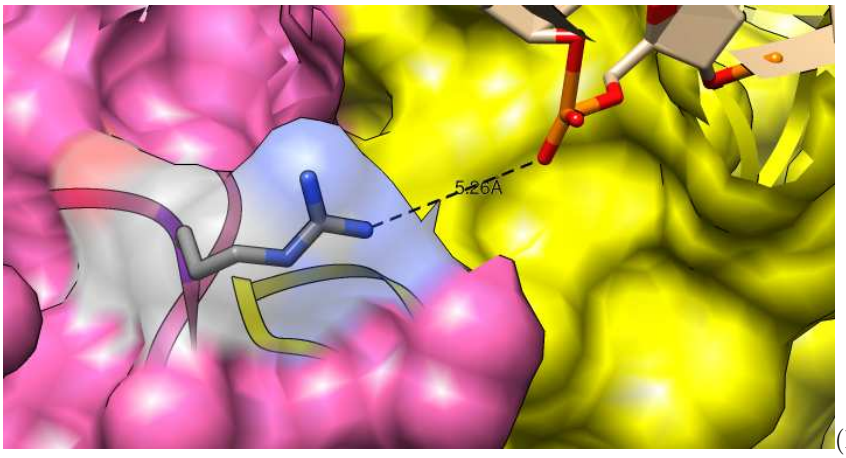


Figure 5 *Topological properties (network centrality measures) highlighted for alignment position 102, which corresponds to mutation Ile195Thr in protein p53 (1TSR.A). (A) Degree heatmap. Degree is relatively conserved as the highlighted column presents a similar shade of orange. (B) Betweenness heatmap. Betweenness is not conserved as column 102 presents many different shades of blue. (C) Closeness heatmap. This measure presents relative conservations in column 102 as we see similar shades of yellow, but there are some positions in this colum that presents very low values for closeness (light sade of yellow).*



(A)



(B)

Figure 6 *Protein-DNA interface in PDB 1TSR. (A) DNA binding overview. Chain A surface is shown in purple and Arg248 is colored by chemical elements. (B) Probable charged attractive interaction between Arg248 and the DNA. Figures generated with Chimera [6].*

Figure 7 *Residue conservation highlighted for alignment position 75, which corresponds to the conservative mutation His168Arg in protein p53 (1TSR.A), and missing residues (166-170), which were experimentally observed for His168Arg mutation (PDB 2BIN), highlighted by the black square.*

Apêndice B

Arquivo suplementar do artigo 2

DETECÇÃO E VISUALIZAÇÃO DE SUBESTRUTURAS FREQUENTES NA INTERFACE PROTEÍNA-LIGANTE EM NÍVEL ATÔMICO ATRAVÉS DE MINERAÇÃO DE SUBGRAFOS FREQUENTES

1 METHODS

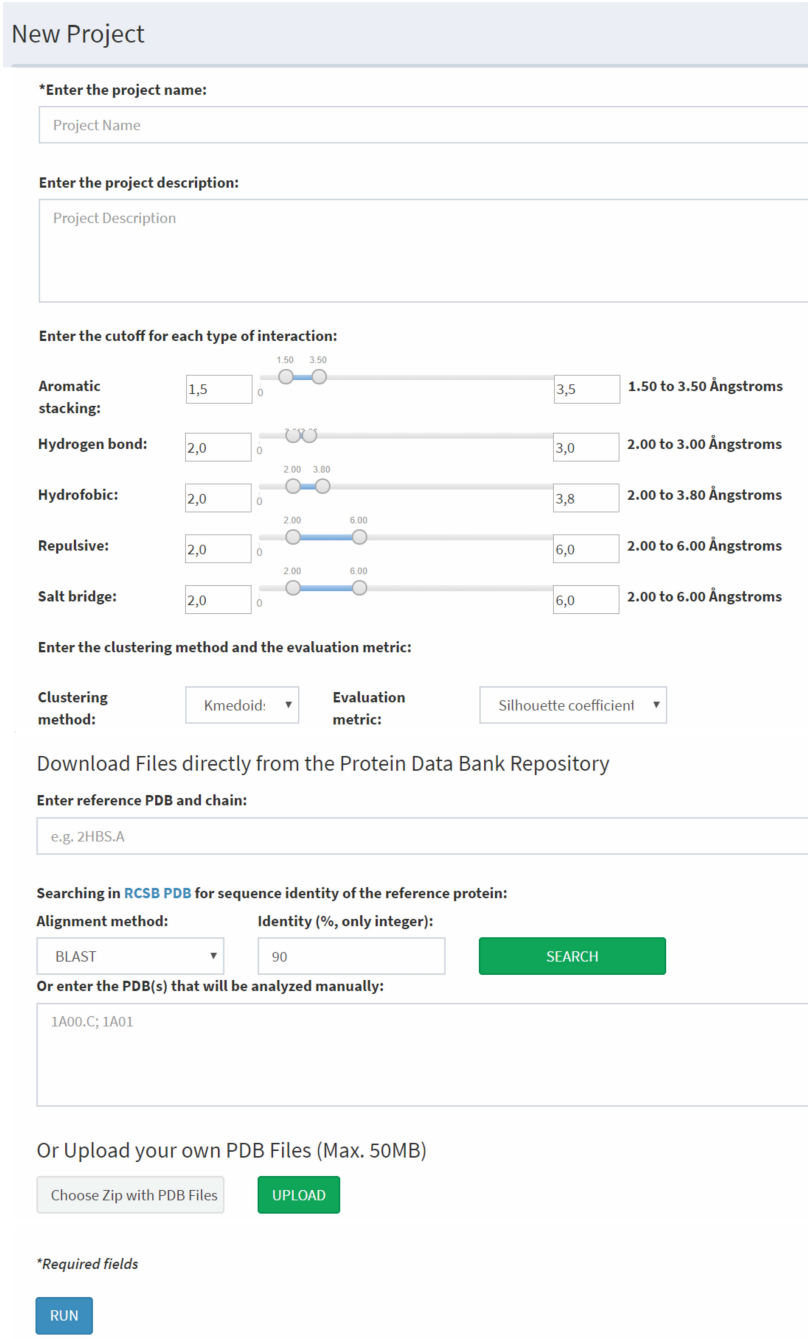


Fig. 1. Starting a new project in visGReMLIN. There are 3 options for the user to provide the dataset of protein-ligand complexes similar to the reference PDB id to be analyzed: **a)** Let our tool to automatically search on PDB for these complexes, by selecting an alignment method and an identity percentage; **b)** Enter a dataset of previously selected complexes manually (type or copy and paste); **c)** Upload users own complexes in PDB format (structures that are not deposited in PDB.

Table 1: Atom types used in visGReMLIN.

Molecule	Atom name	Hydrophobic	Aromatic	Positive	Negative	Donor	Acceptor
ALA	N					X	
ALA	CA						
ALA	C						
ALA	O						X
ALA	CB	X					
ARG	N					X	
ARG	CA						
ARG	C						
ARG	O						X
ARG	CB	X					
ARG	CG	X					
ARG	CD						
ARG	NE			X		X	
ARG	CZ			X			
ARG	NH1			X		X	
ARG	NH2			X		X	
ASN	N					X	
ASN	CA						
ASN	C						
ASN	O						X
ASN	CB	X					
ASN	CG						
ASN	OD1						X
ASN	ND2					X	
ASP	N					X	
ASP	CA						
ASP	C						
ASP	O						X
ASP	CB	X					
ASP	CG						
ASP	OD1				X		X
ASP	OD2				X		X
CYS	N					X	
CYS	CA						
CYS	C						
CYS	O						X
CYS	CB	X					
CYS	SG					X	X
GLN	N					X	
GLN	CA						
GLN	C						
GLN	O						X
GLN	CB	X					
GLN	CG	X					
GLN	CD						
GLN	OE1						X
GLN	NE2					X	
GLU	N					X	
GLU	CA						
GLU	C						
GLU	O						X
GLU	CB	X					
GLU	CG	X					
GLU	CD						
GLU	OE1				X		X

Table 1 continued from previous page

Molecule	Atom name	Hydrophobic	Aromatic	Positive	Negative	Donor	Acceptor
GLU	OE2				X		X
GLY	N					X	
GLY	CA						
GLY	C						
GLY	O						X
HIS	N					X	
HIS	CA						
HIS	C						
HIS	O						X
HIS	CB	X					
HIS	CG		X				
HIS	ND1		X	X		X	X
HIS	CD2		X				
HIS	CE1		X				
HIS	NE2		X	X		X	X
ILE	N					X	
ILE	CA						
ILE	C						
ILE	O						X
ILE	CB	X					
ILE	CG1	X					
ILE	CG2	X					
ILE	CD1	X					
LEU	N					X	
LEU	CA						
LEU	C						
LEU	O						X
LEU	CB	X					
LEU	CG	X					
LEU	CD1	X					
LEU	CD2	X					
LYS	N					X	
LYS	CA						
LYS	C						
LYS	O						X
LYS	CB	X					
LYS	CG	X					
LYS	CD	X					
LYS	CE						
LYS	NZ			X		X	
MET	N					X	
MET	CA						
MET	C						
MET	O						X
MET	CB	X					
MET	CG	X					
MET	SD						X
MET	CE	X					
PHE	N					X	
PHE	CA						
PHE	C						
PHE	O						X
PHE	CB	X					
PHE	CG	X	X				
PHE	CD1	X	X				
PHE	CD2	X	X				

Table 1 continued from previous page

Molecule	Atom name	Hydrophobic	Aromatic	Positive	Negative	Donor	Acceptor
PHE	CE1	X	X				
PHE	CE2	X	X				
PHE	CZ	X	X				
PRO	N						
PRO	CA						
PRO	C						
PRO	O						X
PRO	CB	X					
PRO	CG	X					
PRO	CD						
SER	N					X	
SER	CA						
SER	C						
SER	O						X
SER	CB						
SER	OG					X	X
THR	N					X	
THR	CA						
THR	C						
THR	O						X
THR	CB						
THR	OG1					X	X
THR	CG2	X					
TRP	N					X	
TRP	CA						
TRP	C						
TRP	O						X
TRP	CB	X					
TRP	CG	X	X				
TRP	CD1		X				
TRP	CD2	X	X				
TRP	NE1		X			X	
TRP	CE2		X				
TRP	CE3	X	X				
TRP	CZ2	X	X				
TRP	CZ3	X	X				
TRP	CH2	X	X				
TYR	N					X	
TYR	CA						
TYR	C						
TYR	O						X
TYR	CB	X					
TYR	CG	X	X				
TYR	CD1	X	X				
TYR	CD2	X	X				
TYR	CE1	X	X				
TYR	CE2	X	X				
TYR	CZ		X				
TYR	OH					X	X
VAL	N					X	
VAL	CA						
VAL	C						
VAL	O						X
VAL	CB	X					
VAL	CG1	X					
VAL	CG2	X					

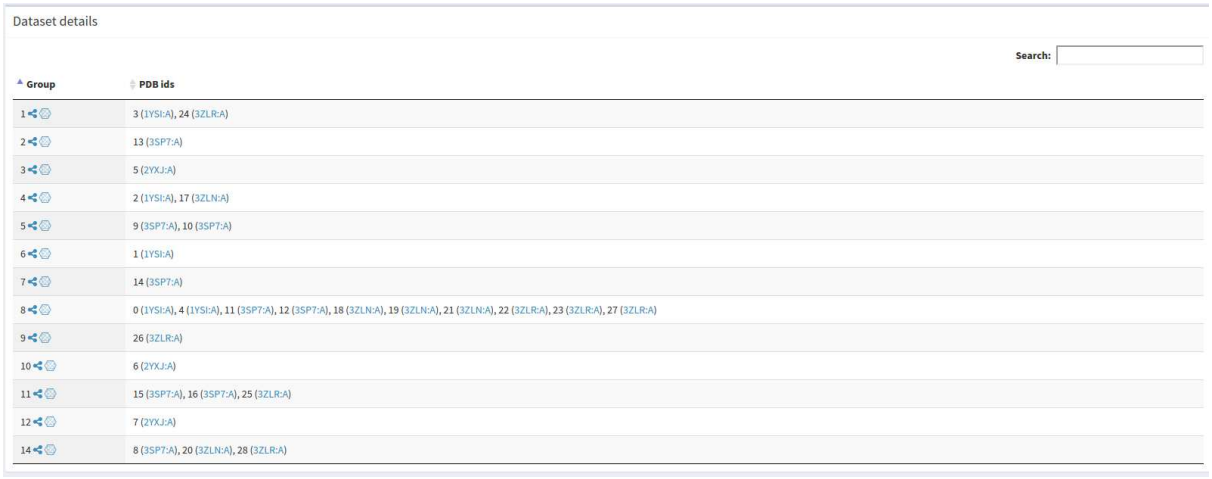


Fig. 2. The Dataset Details section shows the results of the clustering of the connected graphs that represent the interactions between a given protein and ligand.

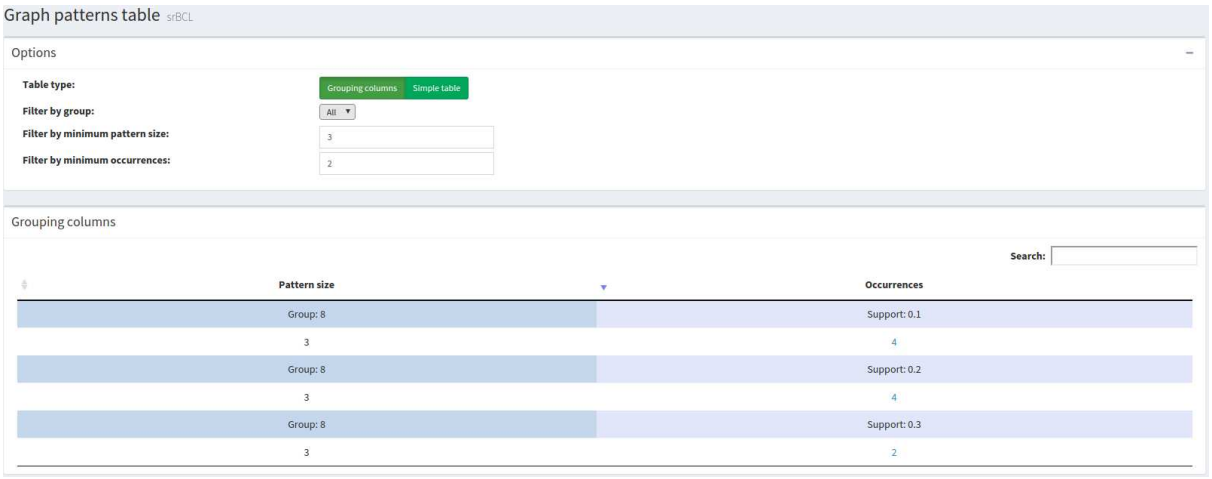


Fig. 3. The Graph patterns table page shows a summary of the patterns found for each group and support.

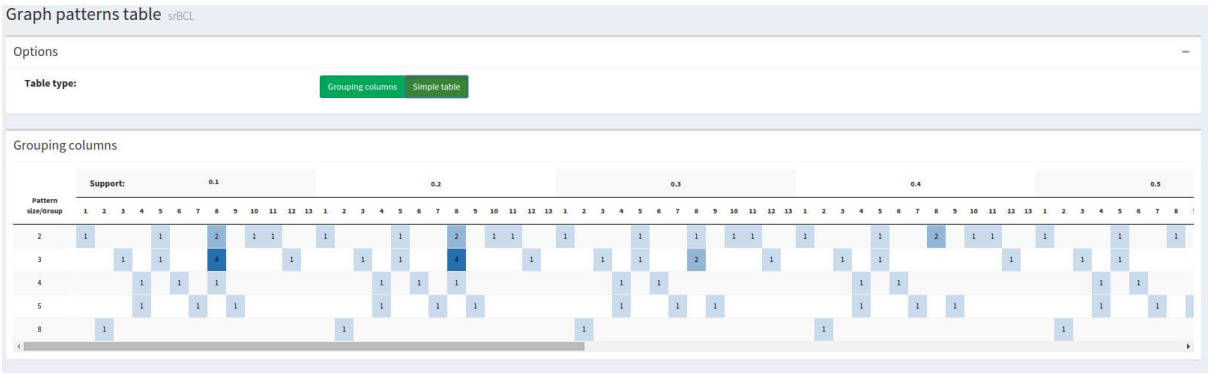


Fig. 4. The simple table shows the number of times that a pattern of a given group, support and size appears.

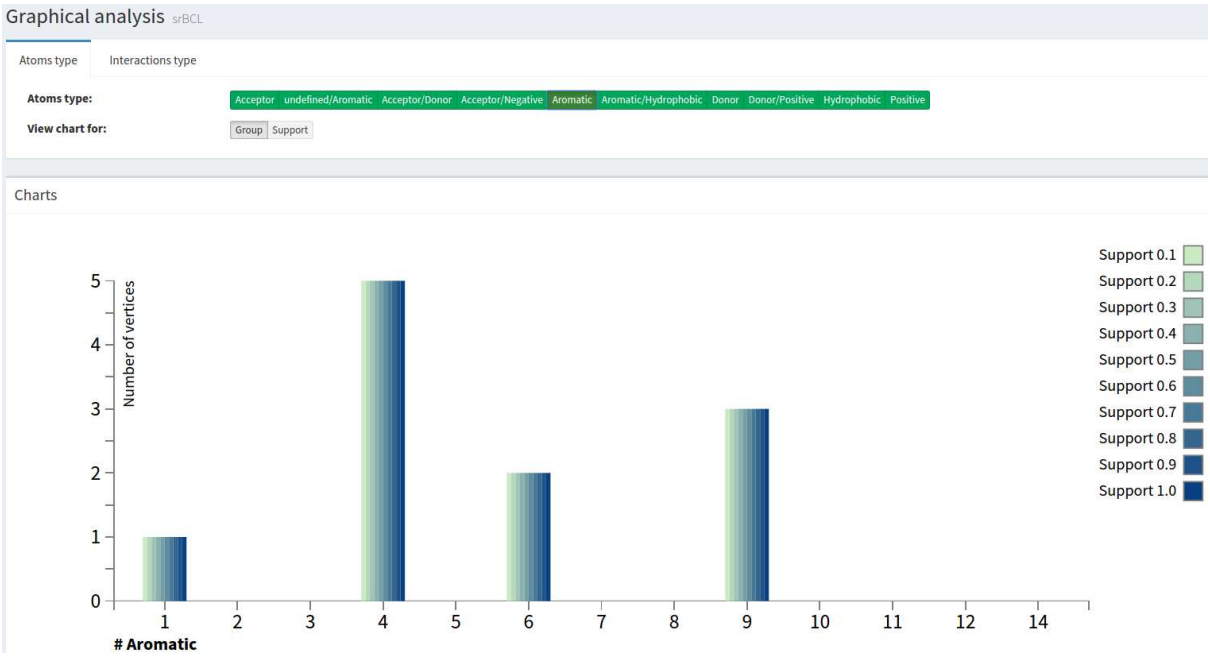


Fig. 5. Summary of motifs extracted from PLI graphs.

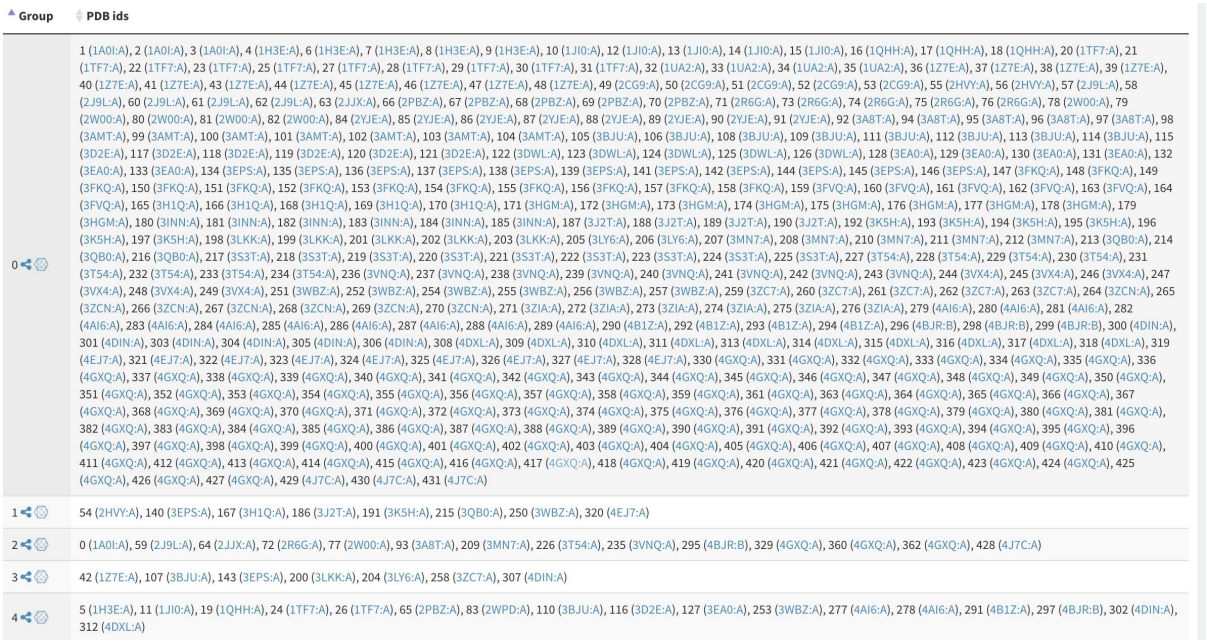
2 RESULTS AND DISCUSSION

Group	PDB ids
0	0 (3QL8:A), 3 (3QQF:A), 6 (3QQG:A), 8 (3QQH:A), 10 (3QQJ:A), 12 (3QQK:A), 14 (3QQK:A), 33 (3QTQ:A), 36 (3QTQ:A), 37 (3QTR:A), 39 (3QTR:A), 41 (3QTS:A), 42 (3QTS:A), 45 (3QTU:A), 47 (3QTU:A), 48 (3QTW:A), 50 (3QTW:A), 52 (3QTX:A), 53 (3QTX:A), 55 (3QTZ:A), 56 (3QTZ:A), 57 (3QU0:A), 58 (3QU0:A), 61 (3QWJ:A), 67 (3QWK:A), 73 (3QX2:A), 77 (3QX4:A), 80 (3QXO:A), 84 (3QXP:A), 85 (3QXP:A), 86 (3QZF:A), 88 (3QZF:A), 91 (3QZG:A), 99 (3QZI:A), 107 (3R1S:A), 109 (3R1Y:A), 116 (3R28:A), 121 (3R6X:A), 124 (3R71:A), 127 (3R73:A), 130 (3R7E:A), 139 (3R7U:A), 142 (3R7V:A), 145 (3R7Y:A), 161 (3R8U:A), 163 (3R8U:A), 166 (3R8V:A), 168 (3R8V:A), 170 (3R8Z:A), 171 (3R8Z:A), 175 (3R9D:A), 176 (3R9D:A), 178 (3R9H:A), 179 (3R9H:A), 184 (3R9N:A), 185 (3R9N:A), 187 (3R9O:A), 188 (3R9O:A), 193 (3RAH:A), 194 (3RAH:A), 199 (3RAI:A), 204 (3RAK:A), 205 (3RAK:A), 207 (3RAL:A), 208 (3RAL:A), 210 (3RJC:A), 211 (3RJC:A), 212 (3RK5:A), 214 (3RK5:A), 215 (3RK7:A), 216 (3RK7:A), 217 (3RK9:A), 219 (3RK9:A), 221 (3RKB:A), 222 (3RKB:A), 231 (3RMF:A), 232 (3RMF:A), 233 (3RNI:A), 234 (3RNI:A), 242 (3RPO:A), 246 (3RPR:A), 248 (3RPR:A), 250 (3RPV:A), 252 (3RPV:A), 254 (3RPV:A), 255 (3RPY:A), 256 (3RPY:A), 259 (3RZB:A), 260 (3RZB:A), 263 (3S00:A), 265 (3S00:A), 266 (3S0O:A), 268 (3S0O:A), 270 (3S1H:A), 271 (3S1H:A), 274 (3SQQ:A), 275 (3SQQ:A)
1	2 (3QQF:A), 5 (3QQF:A), 11 (3QQJ:A), 17 (3QQL:A), 22 (3QRT:A), 24 (3QRT:A), 26 (3QRU:A), 40 (3QTS:A), 49 (3QTW:A), 60 (3QWJ:A), 63 (3QWJ:A), 64 (3QWK:A), 66 (3QWK:A), 68 (3QX2:A), 69 (3QX2:A), 72 (3QX2:A), 74 (3QX4:A), 75 (3QX4:A), 79 (3QXO:A), 82 (3QXO:A), 92 (3QZH:A), 96 (3QZH:A), 98 (3QZI:A), 101 (3QZI:A), 102 (3R1Q:A), 103 (3R1Q:A), 105 (3R1S:A), 106 (3R1S:A), 111 (3R1Y:A), 112 (3R28:A), 113 (3R28:A), 118 (3R6X:A), 119 (3R6X:A), 122 (3R71:A), 123 (3R71:A), 128 (3R73:A), 131 (3R7E:A), 132 (3R7I:A), 133 (3R7I:A), 135 (3R7I:A), 137 (3R7U:A), 138 (3R7U:A), 140 (3R7U:A), 141 (3R7V:A), 144 (3R7Y:A), 148 (3R8L:A), 153 (3R8M:A), 157 (3R8P:A), 160 (3R8U:A), 165 (3R8V:A), 182 (3R9N:A), 191 (3RAH:A), 195 (3RAI:A), 197 (3RAI:A), 201 (3RAI:A), 203 (3RAK:A), 206 (3RAL:A), 213 (3RK5:A), 224 (3RM6:A), 225 (3RM6:A), 226 (3RM6:A), 235 (3ROY:A), 237 (3ROY:A), 239 (3RPO:A), 241 (3RPO:A), 244 (3RPR:A), 253 (3RPV:A), 262 (3S00:A), 272 (3S1H:A), 273 (3SQQ:A)
2	149 (3R8L:A), 159 (3R8P:A), 169 (3R8V:A)
3	4 (3QQF:A), 15 (3QQL:A), 62 (3QWJ:A), 65 (3QWK:A), 70 (3QX2:A), 78 (3QX4:A), 81 (3QXO:A), 94 (3QZH:A), 100 (3QZI:A), 104 (3R1Q:A), 108 (3R1S:A), 110 (3R1Y:A), 117 (3R28:A), 120 (3R6X:A), 125 (3R71:A), 126 (3R73:A), 129 (3R7E:A), 134 (3R7I:A), 136 (3R7U:A), 143 (3R7V:A), 146 (3R7Y:A), 147 (3R83:A), 154 (3R8M:A), 173 (3R8Z:A), 200 (3RAI:A), 227 (3RM6:A), 229 (3RM7:A), 236 (3ROY:A), 238 (3ROY:A), 243 (3RPO:A)
4	1 (3QL8:A), 7 (3QQG:A), 9 (3QQH:A), 23 (3QRT:A), 44 (3QTS:A), 46 (3QTU:A), 51 (3QTX:A), 180 (3R9H:A), 209 (3RAL:A), 230 (3RMF:A), 249 (3RPR:A), 251 (3RPV:A), 257 (3RPY:A)
5	29 (3QRU:A), 115 (3R28:A), 183 (3R9N:A), 192 (3RAH:A), 220 (3RKB:A)
6	13 (3QQK:A), 16 (3QQL:A), 18 (3QQL:A), 20 (3QRT:A), 21 (3QRT:A), 27 (3QRU:A), 28 (3QRU:A), 30 (3QRU:A), 31 (3QRU:A), 32 (3QRU:A), 34 (3QTQ:A), 35 (3QTQ:A), 38 (3QTR:A), 43 (3QTS:A), 59 (3QWJ:A), 71 (3QX2:A), 76 (3QX4:A), 83 (3QXO:A), 89 (3QZG:A), 95 (3QZH:A), 97 (3QZI:A), 114 (3R28:A), 151 (3R8L:A), 152 (3R8L:A), 155 (3R8P:A), 156 (3R8P:A), 164 (3R8U:A), 167 (3R8V:A), 172 (3R8Z:A), 181 (3R9N:A), 189 (3R9O:A), 190 (3RAH:A), 196 (3RAI:A), 198 (3RAI:A), 202 (3RAI:A), 218 (3RK9:A), 223 (3RKB:A), 228 (3RM6:A), 240 (3RPO:A), 247 (3RPR:A), 258 (3RZB:A), 264 (3S00:A), 267 (3S0O:A), 269 (3S0O:A)
7	87 (3QZF:A), 90 (3QZG:A), 245 (3RPR:A)
8	162 (3R8U:A), 174 (3R9D:A), 186 (3R9O:A)
9	25 (3QRT:A), 54 (3QTZ:A), 93 (3QZH:A), 150 (3R8L:A), 158 (3R8P:A), 177 (3R9H:A)
10	19 (3QRT:A), 261 (3S00:A)

Fig. 6. Dataset details for CDK use case. CDK use case resulted in 276 PLI input graphs divided in 11 groups.

Table 2. PDB ids of datasets used - Results and Discussion

Datasets	Number of Complexes	Complexes
ATP	50	1A0I.A, 4EJ7.A, 2R6G.A, 3ZIA.A, 2CG9.A, 3LKK.A, 3HGM.A, 4GXQ.A, 1TF7.A, 2J9L.A, 2W00.A, 3FKQ.A, 3DWL.A, 3FVQ.A, 3VX4.A, 3T54.A, 3EPS.A, 3D2E.A, 3WBZ.A, 3ZCN.A, 3INN.A, 1UA2.A, 2YJE.A, 4BJR.B, 1H3E.A, 4DXL.A, 3J2T.A, 3MN7.A, 2WPD.A, 1JI0.A, 4J7C.A, 4AI6.A, 2PBZ.A, 3K5H.A, 3H1Q.A, 3VNQ.A, 3A8T.A, 1Z7E.A, 3ZC7.A, 4DIN.A, 2HVY.A, 4B1Z.A, 3QB0.A, 2JJX.A, 3S3T.A, 3BJU.A, 3AMT.A, 3EA0.A, 1QHH.A, 3LY6.A
CDK	73	3QL8.A, 3QQF.A, 3QQG.A, 3QQH.A, 3QQJ.A, 3QQK.A, 3QQL.A, 3QRT.A, 3QRU.A, 3QTQ.A, 3QTR.A, 3QTS.A, 3QTU.A, 3QTW.A, 3QTX.A, 3QTZ.A, 3QU0.A, 3QWJ.A, 3QWK.A, 3QX2.A, 3QX4.A, 3QXO.A, 3QXP.A, 3QZF.A, 3QZG.A, 3QZH.A, 3QZI.A, 3R1Q.A, 3R1S.A, 3R1Y.A, 3R28.A, 3R6X.A, 3R71.A, 3R73.A, 3R7E.A, 3R7I.A, 3R7U.A, 3R7V.A, 3R7Y.A, 3R83.A, 3R8L.A, 3R8M.A, 3R8P.A, 3R8U.A, 3R8V.A, 3R8Z.A, 3R9D.A, 3R9H.A, 3R9N.A, 3R9O.A, 3RAH.A, 3RAI.A, 3RAK.A, 3RAL.A, 3RJC.A, 3RK5.A, 3RK7.A, 3RK9.A, 3RKB.A, 3RM6.A, 3RM7.A, 3RMF.A, 3RNI.A, 3ROY.A, 3RPO.A, 3RPR.A, 3RPV.A, 3RPY.A, 3RZB.A, 3S00.A, 3S0O.A, 3S1H.A, 3SQQ.A



Group	PDB ids
0	1 (1A0I:A), 2 (1A0I:A), 3 (1A0I:A), 4 (1H3E:A), 6 (1H3E:A), 7 (1H3E:A), 8 (1H3E:A), 9 (1H3E:A), 10 (1JI0:A), 12 (1JI0:A), 13 (1JI0:A), 14 (1JI0:A), 15 (1JI0:A), 16 (1QHH:A), 17 (1QHH:A), 18 (1QHH:A), 20 (1TF7:A), 21 (1TF7:A), 22 (1TF7:A), 23 (1TF7:A), 25 (1TF7:A), 27 (1TF7:A), 28 (1TF7:A), 29 (1TF7:A), 30 (1TF7:A), 31 (1TF7:A), 32 (1UA2:A), 33 (1UA2:A), 34 (1UA2:A), 35 (1UA2:A), 36 (1Z7E:A), 37 (1Z7E:A), 38 (1Z7E:A), 39 (1Z7E:A), 40 (1Z7E:A), 41 (1Z7E:A), 43 (1Z7E:A), 44 (1Z7E:A), 45 (1Z7E:A), 46 (1Z7E:A), 47 (1Z7E:A), 48 (1Z7E:A), 49 (2CG9:A), 50 (2CG9:A), 51 (2CG9:A), 52 (2CG9:A), 53 (2CG9:A), 55 (2HVY:A), 56 (2HVY:A), 57 (2J9L:A), 58 (2J9L:A), 60 (2J9L:A), 61 (2J9L:A), 62 (2J9L:A), 63 (2JJX:A), 66 (2PBZ:A), 67 (2PBZ:A), 68 (2PBZ:A), 69 (2PBZ:A), 70 (2PBZ:A), 71 (2R6G:A), 73 (2R6G:A), 74 (2R6G:A), 75 (2R6G:A), 76 (2R6G:A), 78 (2W00:A), 79 (2W00:A), 80 (2W00:A), 81 (2W00:A), 82 (2W00:A), 84 (2YJE:A), 85 (2YJE:A), 86 (2YJE:A), 87 (2YJE:A), 88 (2YJE:A), 89 (2YJE:A), 90 (2YJE:A), 91 (2YJE:A), 92 (3A8T:A), 94 (3A8T:A), 95 (3A8T:A), 96 (3A8T:A), 97 (3A8T:A), 98 (3AMT:A), 99 (3AMT:A), 100 (3AMT:A), 101 (3AMT:A), 102 (3AMT:A), 103 (3AMT:A), 104 (3AMT:A), 105 (3BJU:A), 106 (3BJU:A), 108 (3BJU:A), 109 (3BJU:A), 111 (3BJU:A), 112 (3BJU:A), 113 (3BJU:A), 114 (3BJU:A), 115 (3D2E:A), 117 (3D2E:A), 118 (3D2E:A), 119 (3D2E:A), 120 (3D2E:A), 121 (3D2E:A), 122 (3DWL:A), 123 (3DWL:A), 124 (3DWL:A), 125 (3DWL:A), 126 (3DWL:A), 128 (3EA0:A), 129 (3EA0:A), 130 (3EA0:A), 131 (3EA0:A), 132 (3EA0:A), 133 (3EA0:A), 134 (3EPS:A), 135 (3EPS:A), 136 (3EPS:A), 137 (3EPS:A), 138 (3EPS:A), 139 (3EPS:A), 141 (3EPS:A), 142 (3EPS:A), 144 (3EPS:A), 145 (3EPS:A), 146 (3EPS:A), 147 (3FKQ:A), 148 (3FKQ:A), 149 (3FKQ:A), 150 (3FKQ:A), 151 (3FKQ:A), 152 (3FKQ:A), 153 (3FKQ:A), 154 (3FKQ:A), 155 (3FKQ:A), 156 (3FKQ:A), 157 (3FKQ:A), 158 (3FKQ:A), 159 (3FVQ:A), 160 (3FVQ:A), 161 (3FVQ:A), 162 (3FVQ:A), 163 (3FVQ:A), 164 (3FVQ:A), 165 (3H1Q:A), 166 (3H1Q:A), 168 (3H1Q:A), 169 (3H1Q:A), 170 (3H1Q:A), 171 (3HGM:A), 172 (3HGM:A), 173 (3HGM:A), 174 (3HGM:A), 175 (3HGM:A), 176 (3HGM:A), 177 (3HGM:A), 178 (3HGM:A), 179 (3HGM:A), 180 (3INN:A), 181 (3INN:A), 182 (3INN:A), 183 (3INN:A), 184 (3INN:A), 185 (3INN:A), 187 (3J2T:A), 188 (3J2T:A), 189 (3J2T:A), 190 (3J2T:A), 192 (3K5H:A), 193 (3K5H:A), 194 (3K5H:A), 195 (3K5H:A), 196 (3K5H:A), 197 (3K5H:A), 198 (3LKK:A), 199 (3LKK:A), 201 (3LKK:A), 202 (3LKK:A), 203 (3LKK:A), 205 (3LY6:A), 206 (3LY6:A), 207 (3MN7:A), 208 (3MN7:A), 210 (3MN7:A), 211 (3MN7:A), 212 (3MN7:A), 213 (3QB0:A), 214 (3QB0:A), 216 (3QB0:A), 217 (3S3T:A), 218 (3S3T:A), 219 (3S3T:A), 220 (3S3T:A), 221 (3S3T:A), 222 (3S3T:A), 223 (3S3T:A), 224 (3S3T:A), 225 (3S3T:A), 227 (3T54:A), 228 (3T54:A), 229 (3T54:A), 230 (3T54:A), 231 (3T54:A), 232 (3T54:A), 233 (3T54:A), 234 (3T54:A), 236 (3VNQ:A), 237 (3VNQ:A), 238 (3VNQ:A), 239 (3VNQ:A), 240 (3VNQ:A), 241 (3VNQ:A), 242 (3VNQ:A), 243 (3VNQ:A), 244 (3VX4:A), 245 (3VX4:A), 246 (3VX4:A), 247 (3VX4:A), 248 (3VX4:A), 249 (3VX4:A), 251 (3WBZ:A), 252 (3WBZ:A), 254 (3WBZ:A), 255 (3WBZ:A), 256 (3WBZ:A), 257 (3WBZ:A), 259 (3ZC7:A), 260 (3ZC7:A), 261 (3ZC7:A), 262 (3ZC7:A), 263 (3ZC7:A), 264 (3ZCN:A), 265 (3ZCN:A), 266 (3ZCN:A), 267 (3ZCN:A), 268 (3ZCN:A), 269 (3ZCN:A), 270 (3ZCN:A), 271 (3ZIA:A), 272 (3ZIA:A), 273 (3ZIA:A), 274 (3ZIA:A), 275 (3ZIA:A), 276 (3ZIA:A), 279 (4AI6:A), 280 (4AI6:A), 281 (4AI6:A), 282 (4AI6:A), 283 (4AI6:A), 284 (4AI6:A), 285 (4AI6:A), 286 (4AI6:A), 287 (4AI6:A), 288 (4AI6:A), 289 (4AI6:A), 290 (4B1Z:A), 292 (4B1Z:A), 293 (4B1Z:A), 294 (4B1Z:A), 296 (4BJR:B), 298 (4BJR:B), 299 (4BJR:B), 300 (4DIN:A), 301 (4DIN:A), 303 (4DIN:A), 304 (4DIN:A), 305 (4DIN:A), 306 (4DIN:A), 308 (4DXL:A), 309 (4DXL:A), 310 (4DXL:A), 311 (4DXL:A), 313 (4DXL:A), 314 (4DXL:A), 315 (4DXL:A), 316 (4DXL:A), 317 (4DXL:A), 318 (4DXL:A), 319 (4EJ7:A), 321 (4EJ7:A), 322 (4EJ7:A), 323 (4EJ7:A), 324 (4EJ7:A), 325 (4EJ7:A), 326 (4EJ7:A), 327 (4EJ7:A), 328 (4EJ7:A), 330 (4GXQ:A), 331 (4GXQ:A), 332 (4GXQ:A), 333 (4GXQ:A), 334 (4GXQ:A), 335 (4GXQ:A), 336 (4GXQ:A), 337 (4GXQ:A), 338 (4GXQ:A), 339 (4GXQ:A), 340 (4GXQ:A), 341 (4GXQ:A), 342 (4GXQ:A), 343 (4GXQ:A), 344 (4GXQ:A), 345 (4GXQ:A), 346 (4GXQ:A), 347 (4GXQ:A), 348 (4GXQ:A), 349 (4GXQ:A), 350 (4GXQ:A), 351 (4GXQ:A), 352 (4GXQ:A), 353 (4GXQ:A), 354 (4GXQ:A), 355 (4GXQ:A), 356 (4GXQ:A), 357 (4GXQ:A), 358 (4GXQ:A), 359 (4GXQ:A), 361 (4GXQ:A), 363 (4GXQ:A), 364 (4GXQ:A), 365 (4GXQ:A), 366 (4GXQ:A), 367 (4GXQ:A), 368 (4GXQ:A), 369 (4GXQ:A), 370 (4GXQ:A), 371 (4GXQ:A), 372 (4GXQ:A), 373 (4GXQ:A), 374 (4GXQ:A), 375 (4GXQ:A), 376 (4GXQ:A), 377 (4GXQ:A), 378 (4GXQ:A), 379 (4GXQ:A), 380 (4GXQ:A), 381 (4GXQ:A), 382 (4GXQ:A), 383 (4GXQ:A), 384 (4GXQ:A), 385 (4GXQ:A), 386 (4GXQ:A), 387 (4GXQ:A), 388 (4GXQ:A), 389 (4GXQ:A), 390 (4GXQ:A), 391 (4GXQ:A), 392 (4GXQ:A), 393 (4GXQ:A), 394 (4GXQ:A), 395 (4GXQ:A), 396 (4GXQ:A), 397 (4GXQ:A), 398 (4GXQ:A), 399 (4GXQ:A), 400 (4GXQ:A), 401 (4GXQ:A), 402 (4GXQ:A), 403 (4GXQ:A), 404 (4GXQ:A), 405 (4GXQ:A), 406 (4GXQ:A), 407 (4GXQ:A), 408 (4GXQ:A), 409 (4GXQ:A), 410 (4GXQ:A), 411 (4GXQ:A), 412 (4GXQ:A), 413 (4GXQ:A), 414 (4GXQ:A), 415 (4GXQ:A), 416 (4GXQ:A), 417 (4GXQ:A), 418 (4GXQ:A), 419 (4GXQ:A), 420 (4GXQ:A), 421 (4GXQ:A), 422 (4GXQ:A), 423 (4GXQ:A), 424 (4GXQ:A), 425 (4GXQ:A), 426 (4GXQ:A), 427 (4GXQ:A), 429 (4J7C:A), 430 (4J7C:A), 431 (4J7C:A)
1	54 (2HVY:A), 140 (3EPS:A), 167 (3H1Q:A), 186 (3J2T:A), 191 (3K5H:A), 215 (3QB0:A), 250 (3WBZ:A), 320 (4EJ7:A)
2	0 (1A0I:A), 59 (2J9L:A), 64 (2JJX:A), 72 (2R6G:A), 77 (2W00:A), 93 (3A8T:A), 209 (3MN7:A), 226 (3T54:A), 235 (3VNQ:A), 295 (4BJR:B), 329 (4GXQ:A), 360 (4GXQ:A), 362 (4GXQ:A), 428 (4J7C:A)
3	42 (1Z7E:A), 107 (3BJU:A), 143 (3EPS:A), 200 (3LKK:A), 204 (3LY6:A), 258 (3ZC7:A), 307 (4DIN:A)
4	5 (1H3E:A), 11 (1JI0:A), 19 (1QHH:A), 24 (1TF7:A), 26 (1TF7:A), 65 (2PBZ:A), 83 (2WPD:A), 110 (3BJU:A), 116 (3D2E:A), 127 (3EA0:A), 253 (3WBZ:A), 277 (4AI6:A), 278 (4AI6:A), 291 (4B1Z:A), 297 (4BJR:B), 302 (4DIN:A), 312 (4DXL:A)

Fig. 7. Dataset details for ATP use case. ATP use case resulted in 432 PLI input graphs divided in 5 groups.

Apêndice C

Trabalhos publicados ou em análise para publicação

A seguir estão relacionados os trabalhos publicados ou em análise para publicação, resultantes desta dissertação:

- Apresentação de pôster
  Título: visGReMLIN: An interactive strategy to visualize common substructures in protein-ligand interaction
  Autores: Vagner Soares Ribeiro, Charles A. Santana, Fabio Ribeiro Cerqueira, Alexandre Victor Fassio, Carlos H. Da Silveira, Raquel Melo Minardi and Sabrina De Azevedo Silveira
  Evento: X-Meeting 2016 - 12th International Conference of the AB3C
  Local: Belo Horizonte - Brasil.
  Ano: 2016.

- Artigo publicado
  Título: Vermont: a multi-perspective visual interactive platform for mutational analysis
  Autores: Alexandre V Fassio, Pedro M Martins, Samuel da S Guimarães, Sócrates S A Junior, Vagner S Ribeiro, Raquel C de Melo-Minardi e Sabrina de A Silveira

Revista: BMC Bioinformatics
Ano: 2017.

- Trabalho enviado para publicação (aguardando resposta)
  Título: Detecção e visualização de subestruturas comuns na interface proteína-ligante em nível atômico através de mineração de subgrafos frequentes
  Autores: Vagner S. Ribeiro, Charles A. Santana, Alexandre V. Fassio, Adriana M. Patarroyo-Vargas, Maria G. A. Oliveira, Valdete M. Gonçalves-Almeida, Sandro C. Izidoro, Raquel C. de Melo-Minardi e Sabrina de A. Silveira
  Evento: ISMB (*International Conference on Intelligent Systems for Molecular Biology*) 2018
  Local: Chicago - Estados Unidos.
  Ano: 2018.