Bioinformatics, 31(18), 2015, 3027–3034
doi: 10.1093/bioinformatics/btv274
Advance Access Publication Date: 15 May 2015
Original paper

OXFORD

Databases and ontologies

SpeeDB: fast structural protein searches

David E. Robillard[1], Phelelani T. Mpangase[2], Scott Hazelhurst[2,3] and Frank Dehne[1]

[1]School of Computer Science, Carleton University, Ottawa, Ontario, Canada, [2]Sydney Brenner Institute for Molecular Bioscience and [3]School of Electrical and Information Engineering, University of the Witwatersrand, Johannesburg, South Africa

Associate Editor: Janet Kelso

Received on February 26, 2014; revised on April 25, 2015; accepted on April 27, 2015

Abstract

Motivation: Interactions between amino acids are important determinants of the structure, stability and function of proteins. Several tools have been developed for the identification and analysis of such interactions in proteins based on the extensive studies carried out on high-resolution structures from Protein Data Bank (PDB). Although these tools allow users to identify and analyze interactions, analysis can only be performed on one structure at a time. This makes it difficult and time consuming to study the significance of these interactions on a large scale.

Results: SpeeDB is a web-based tool for the identification of protein structures based on structural properties. SpeeDB queries are executed on all structures in the PDB at once, quickly enough for interactive use. SpeeDB includes standard queries based on published criteria for identifying various structures: disulphide bonds, catalytic triads and aromatic–aromatic, sulphur–aromatic, cation–π and ionic interactions. Users can also construct custom queries in the user interface without any programming. Results can be downloaded in a Comma Separated Value (CSV) format for further analysis with other tools. Case studies presented in this article demonstrate how SpeeDB can be used to answer various biological questions. Analysis of human proteases revealed that disulphide bonds are the predominant type of interaction and are located close to the active site, where they promote substrate specificity. When comparing the two homologous G protein-coupled receptors and the two protein kinase paralogs analyzed, the differences in the types of interactions responsible for stability accounts for the differences in specificity and functionality of the structures.

Availability and implementation: SpeeDB is available at http://www.parallelcomputing.ca as a web service.

Contact: d@drobilla.net

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The majority of biological processes in living organisms are carried out by proteins. Proteins can function as enzymes, which catalyze and regulate chemical reactions. Some proteins transport and store essential biological molecules such as oxygen, metal ions and glucose. Other proteins are found in connective tissues and function as structural elements; collagen is an example of a structural protein which occurs in all multicellular organisms and functions as the major stress-bearing component of connective tissues. The function of proteins is determined by the correct folding of the amino acid sequence into a stable three-dimensional structure (Goldenberg, 1985; Sundaramurthy *et al.*, 2010). With so many different roles, it is crucial that proteins are able to recognize and interact with their specific substrates in order to carry out their function (Pitre *et al.*, 2006).

Interactions that occur between amino acid residues are the main factors that determine the structure, stability and function of proteins (Chourasia *et al.*, 2011; Tina *et al.*, 2007). A number of

interactions have been associated with protein stability and assembly into their native structures. These include weak hydrogen bonds, ionic interactions, disulphide bridges, aromatic interactions and hydrophobic interactions (Tina *et al.*, 2007). Being able to determine which interactions occur in which proteins has many applications in characterizing protein function, including drug design. For example, Cerutti *et al.* (2010) had to find a candidate protein with specific disulphide bonds as a key step in HIV drug design. The ability to support similar future work was a key motivation for SpeeDB.

A key source of protein interactions is the protein data bank (PDB), which currently has over 80 000 resolved protein structures, with the relative 3D positions of each atom determined using experimental methods such as X-ray crystallography. Interactions are either explicitly annotated by the crystallographers or can be inferred by the proximity and configuration of amino acids, together with the amino acid or atom properties.

The goal of this work is to present a tool that allows the entire PDB to be searched for a wide variety of user-specified interactions, rapidly enough for interactive use.

1.1 Related work

A number of tools have been developed for the recognition and analysis of the different types of interactions in proteins. These tools are based on results from extensive studies on individual interactions, carried out on high-resolution protein structures from the PDB (Berman *et al.*, 2000), and aim to understand how interactions are distributed in proteins and the geometric parameters for identifying them. These tools include the protein interactions calculator (PIC) server (Tina *et al.*, 2007), which calculates different types of interactions that occur within a given protein structure, the aromatic–aromatic Interactions Database (A2ID) (Chourasia *et al.*, 2011) which focuses mainly on the aromatic–aromatic interactions within proteins, and IntGeom (Pal *et al.*, 2009) which calculates the interaction geometry between planar groups in proteins.

However, though the available tools allow users to identify and analyze different types of interactions found in proteins, analysis can only be done by submitting a single structure at a time. This makes it difficult to study the significance of the role played by the different interactions in proteins on a large scale, or to identify unknown proteins in which the interaction(s) of interest occur.

Earlier work of ours, PH2 (Hazelhurst, 2010) relied on massive parallelism to achieve a limited range of the functionality of SpeeDB. SpeeDB is at least an order of magnitude faster, several orders of magnitudes more efficient, supports more advanced queries and provides an interactive interface.

1.2 Summary of contributions

In this study, we present SpeeDB, a web-based tool for the identification of crystal structures of proteins from the PDB based on the different types of amino acid interactions that occur within the folded protein. Interaction queries have been built based on published criteria for identifying such interactions.

SpeeDB queries are based on pairs of atoms with distance constraints, e.g. 'a sulphur on a cysteine that is between 0.1 and 2.2 Å from another sulphur on a cysteine'. A query may consist of many such pairs grouped into patterns which correspond to amino acids, as well as annotated connections provided by the crystallographer such as covalent bonds. Searches can be restricted to specific groups of proteins using filters such as organism, resolution and SCOP classification.

The main contribution of this article is the performance of SpeeDB compared to that of previous systems. SpeeDB's query functionality is comparable to the PIC server (Tina *et al.*, 2007). However, for a given query, PIC can only search a single PDB structure at a time. As demonstrated in Section 4.3, there are various use cases where it is important to query all (over 80 000) structures in the PDB. This is practically impossible with PIC. Earlier work of ours, PH2 (Hazelhurst, 2010), relied on massive parallelism to address this problem. For queries with limited functionality (compared to PIC and SpeeDB), PH2 can search the entire PDB in about 4 minutes on an 838 core cluster, or about 100 minutes on a 40 core cluster. SpeeDB provides more advanced functionality (including all the functionality of PIC), and allows querying all structures in the PDB in a few seconds, on only one standard desktop PC. SpeeDB is the first system fast enough to support interactive use, allowing investigators to experiment with different query parameters in a real-time fashion.

The technical innovations introduced by SpeeDB include a novel in-memory database structure for PDB files and a multithreaded query server. Previous systems like PIC and PH2 relied on search algorithms with quadratic query time. SpeeDB instead uses a spatial data structure and query mechanism that allows querying PDB structures in sublinear time. In conjunction with a highly optimized memory layout and multithreaded query engine, this allows SpeeDB to achieve query performance several orders of magnitude faster than previous work.

Another important contribution of SpeeDB is a user-friendly and powerful user interface that allows users to build up flexible queries without programming.

The case studies presented here demonstrate the uniqueness of SpeeDB and the extent to which the resource can be used to answer various biological questions. Using SpeeDB, we were able to analyze the distribution of interactions within proteins, identify the predominant interactions in different families of proteins and identify important interactions known to be responsible for structural stability.

2 Background: interactions in proteins

There are different types of interactions in proteins. Some of these interactions are annotated in the PDB data. Other interactions can be detected by using spatial and biochemical properties of the atoms involved. The PDB records contain 3D positions of atoms for each amino acid in the protein. By using these positions, it is possible to identify the different types of interactions by calculating the distances and geometry between the atoms of different amino acids. The core of SpeeDB is an algorithm which very rapidly allows for the identification of interactions using the positions and other properties of atoms found in the PDB records. For the algorithm to successfully identify these interactions, certain criteria and parameters must be met by the atoms thought to interact. SpeeDB allows users to specify interactions through 3D and other properties in a very general way. In addition, it supports some key interactions as packaged queries—these are described below.

2.1 Hydrogen bonds

Hydrogen bonds are the most common types of interactions in proteins. They are well known for stabilizing protein structures, protein–protein interfaces as well as protein–ligand complexes (Manikandan and Ramakumar, 2004). Hydrogen bonds form between weak acidic groups (hydrogen donor) and weak bases (hydrogen acceptor). The donor group can be N–H or O–H, while the

hydrogen acceptor can be an N or O with a lone pair of electrons. Most hydrogen bonds in proteins are local (between amino acids that are close to each other in a peptide chain) and occur between the atoms of the polypeptide backbone, where their role is to stabilize the secondary structures of proteins (Stickle *et al.*, 1992). The α-helices and β-sheets in proteins are stabilized by hydrogen bonds that occur between the amide hydrogen (N–H) and carboxyl oxygen (C = O) of the polypeptide backbone.

2.2 Ionic interactions

The positively (histidine, arginine and lysine) and negatively (aspartic acid and glutamic acid) charged amino acids have the potential to form both repulsive and attractive electrostatic interactions in proteins as a result of their charges. The ionic interactions that form between the oppositely charged amino acid residues (ion pairs) have been studied by Barlow and Thornton (1983). In their study, the ion pairs were analyzed in 38 different crystal structures of proteins for their contribution towards stabilizing the protein structure. They determined that the ionpairs are responsible for stabilizing the tertiary structure since 76% of the observed ionpairs link together different elements of the secondary structure (α-helices and β-sheets).

2.3 Disulphide bonds

The disulphide bonds that are formed between the sulphur atoms of two cysteine residues have been implicated in both folding and stability of native protein structures. Site-directed mutagenesis experiments have been performed to introduce disulphide bonds in proteins in order to increase the stability of their native conformations by limiting the mobility of portions of the polypeptide chain (Fersht and Serrano, 1993; Goldenberg, 1985). Sowdhamini *et al.* (1989) designed a model for identifying the optimal positions in proteins for disulphide bond introduction. This model was tested on crystal structures of proteins in which mutant disulphide bonds had been introduced. In their analysis, they identified that in all these proteins, the introduction of the disulphide bonds did not contribute towards stability. This was because the mutant disulphide bonds were introduced at sites where they were 'steriochemically non-optimal' or 'strained'. It is thus crucial for site-directed mutagenesis studies to introduce disulphide bonds at positions that allow for correct cross-links in order to increase the stability of proteins. This method by Sowdhamini *et al.* (1989) not only allows identification of sites for introduction of disulphide bonds, but also allows identification of the steriochemically correct disulphide bonds in proteins.

2.4 Aromatic interactions

Amino acids phenylalanine, tyrosine and tryptophan are often found in globular proteins where they are buried in the interior, near nonpolar amino acids (Burley and Petsko, 1985). These aromatic amino acids play an important role in protein stability, ligand recognition, DNA recognition and protein–protein interactions (Babu, 2003; Chourasia *et al.*, 2011; Meyer *et al.*, 2003; Tina *et al.*, 2007). The aromatic rings of these amino acids carry partial positive charges (δ+) on their edges and partial negative charge (δ−) on their surfaces due to the delocalization of the π-electrons. This allows aromatic rings to interact with each other and other amino acids through electrostatic forces (Scrutton and Raine, 1996).

The analysis of high-resolution crystal structures of proteins from the PDB has revealed three main types of interactions that stabilize protein structures which are associated with aromatic amino acids. These are aromatic–aromatic, sulphur–aromatic and cation–π interactions. Burley and Petsko (1985) analyzed a total of 34 crystal structures of proteins for aromatic–aromatic interactions and identified that on average, 61% of phenylalanine, 54% of tyrosine and 59% of tryptophan residues are involved in aromatic–aromatic interactions. They also found that 80% of the aromatic–aromatic interactions identified link unique secondary structural elements (α-helices and β-sheets) and that most of these interactions are energetically favourable, which contributes to the protein's tertiary structure.

Sulphur–aromatic interactions occur between the sulphur-containing amino acids (cysteine and methionine) and aromatic amino acids. The sulphur–aromatic interactions were studied by Reid *et al.* (1985) using 36 high-resolution crystal structures of proteins. In their study, they identified that almost half of the sulphur atoms from cysteine (S–H) and more than half the sulphur atoms from cysteine (S–S) and methionine are within 6Å from the aromatic ring centroid and predominantly approached the δ+ edge of the aromatic rings as compared to the δ–π-cloud. Their data also showed that of the interacting sulphur-containing residues, many are involved in an interaction with more than one aromatic residues.

The cation–π, or rather 'amino–aromatic', interactions occur between the side chains of the positively charged (or δ+) amino acids (arginine and lysine) and the δ–π-system of the aromatic amino acids. Burley and Petsko (1986) studied the geometry and frequency of the interactions between the amino groups of the cationic amino acids and aromatic amino acids by analyzing 33 high-resolution crystal structures of proteins. They showed that ∼50% of each of the different aromatic rings interacts with the amino groups. They also showed that 50 and 25% of the arginine and lysine residues observed interact with the aromatic rings, respectively. Analysis of the geometry of the interaction revealed that the positively charged or δ+ amino groups are found adjacent to the δ–π-electron cloud of the aromatic ring, away from the δ+ edge of the aromatic ring. The interaction distance observed between the aromatic ring centroids and the nitrogen atoms of the amino group was between 3.4 and 6 Å.

3 Methods

3.1 Database design

SpeeDB is a bespoke in-memory database with a compact flat structure. Compact data along with careful attention to low-level details such as efficient memory allocation allow the entire PDB data set to be scanned quickly enough for interactive use.

The main section of the database is a sequence of protein structures. Each protein begins with a header of basic information (e.g. ID, resolution) followed by three sequences: entities (e.g. polymer), annotated connections (e.g. covalent bond), and models. A model is a sequence of atoms, each of which has identifying information (e.g. chain, atom type) along with a 3D geometric coordinate. All data for one protein is packed in a single contiguous chunk of memory, with one exception: long strings like experiment titles are stored in a separate section in order to compact the crucial information as much as possible.

To construct the database, a separate program loads the required information from the mmCIF (Westbrook and Bourne, 2000) files provided by the PDB project, and writes the database to a single file in the machine's native binary format. This process is very time intensive, but only needs to be performed once. The web server application then maps this file into memory on start-up, which takes very little time.

In this format, the complete PDB data set of roughly 80 000 proteins fits in under 15 GB, which can be stored entirely in RAM on current PC hardware. Thus, given sufficient memory SpeeDB is effectively an in-memory database, but due to the use of memory mapping, the system will also function correctly on machines with less memory.

3.2 Query algorithm

The fundamental task of the SpeeDB engine is to search the database as quickly as possible for proteins that match a given query. A query consists of three types of information:

- Protein filters (e.g. resolution)
- Connections, each of which contains a type (e.g. covalent bond) and minimum/maximum distance constraints. These match connection annotations given in the PDB data.
- Patterns (distance constraints), each of which contains a *root* amino acid, and one or more *branches* from the root amino acid. Each branch has the atom type of the *root atom*, the amino acid and atom type of the *leaf atom*, a chain constraint, and minimum/maximum distance constraints. A pattern only matches if a match is found for all branches, i.e. a pattern is an AND of its branches.

The query algorithm scans over the entire database of proteins, checking if each matches the query and reporting results if so. For performance, the objective is to reject nonmatching proteins as quickly as possible.

For each protein, the matching algorithm is as follows:

1. Check all protein filters. If any do not match, reject this protein.
2. For each query connection, check if the protein contains a connection annotation with matching type and distance. If not, reject this protein.
3. For each model:
   a. For each pattern:
      1. Scan the model to find all atoms that match a root or leaf, and store these in sets of temporary arrays R and L, respectively. R only contains matches for amino acids which contain a matching root atom for every branch. L contains one array of atom pointers for each branch in the pattern.
      2. Build one KD-Tree (Bentley, 1975) per branch with atoms in L.
      3. For each branch, for each potential root in R: use the corresponding tree to find atoms within the required distance. If such an atom is found, record the match and continue.
      4. If there is not at least one match for each branch, reject the protein.
4. If no atom match is found in any model, reject the protein.

3.3 Annotated connections

The PDB data often includes a category that describes interactions between different parts of the protein structure. The interaction partners are identified by the chain, position, and amino acid, as well as the individual atoms. This information is included in the SpeeDB database, so queries can specify 'Annotated Connections' that are found in the protein structure files. These include covalent bonds, disulphide bonds and metal coordinations. For example, a query can match only structures that include annotated covalent bonds with distance within 1.5–2 Å.

3.4 Structures with multiple models

For structures with more than one model (e.g. such as Nuclear Magnetic Resonance (NMR) structures), SpeeDB examines all models. The search results in the web interface show all PDB entries with a match in any model. The investigator can then choose to see the details for a given entry, where the matches for all models are shown. This information is also available in the CSV export, and thus can be analyzed in any way the investigator desires. Future work could include a facility to report only matches with a given model constraint, for example, only structures where the majority or all of the models match the query.

3.5 Interaction queries methodology

3.5.1 Disulphide bonds

The disulphide bond query was built based on the analysis done by (Sowdhamini *et al.*, 1989). If two sulphur atoms from the cysteine residues are within 2.2Å from each other, this structure qualifies as a disulphide bond. The steriochemistry of the disulphide bonds were not taken into consideration as these were beyond the scope of this project.

3.5.2 Aromatic–aromatic interactions

If the centroids of aromatic amino acids are within 4.5–7Å of each other, this structure qualifies as an aromatic–aromatic interaction (Burley and Petsko, 1985; Chourasia *et al.*, 2011).

3.5.3 Sulphur–aromatic interactions

Aromatic–sulphur interactions occur between the sulphur-containing amino acids (cysteine and methionine) and the aromatic ring. If the sulphur atom of cysteine or methionine is within 5.3Å of the aromatic ring's centroid, this structure qualifies as a sulphur–aromatic interaction (Reid *et al.*, 1985).

3.5.4 Cation–π/amino–aromatic interactions

Cation–π interactions are formed between the positively charged amino acids (lysine and arginine) and aromatic rings. If any of the nitrogen atoms from the amino groups of lysine (NZ) and arginine (NH1 or NH2) are within 6Å of the centroid of the aromatic rings of phenylalanine, tyrosine or tryptophan, then the interaction qualifies as a cation–π/amino–aromatic interaction (Sathyapriya and Vishveshwara, 2004).

3.5.5 Ionic interactions

To identify the ionic interactions between amino acids, the distance between the atoms of the charged groups of the side chains in the positively (histidine, arginine and lysine) and negatively (aspartic acid and glutamic acid) charged amino acids were considered. If any of the nitrogen atoms from the positively charged amino group are within a distance of 6Å from the oxygen atom of the negatively charged carboxyl group, the interaction qualifies as an ionic interaction (Mihel *et al.*, 2008; Tina *et al.*, 2007).

3.6 Implementation details

SpeeDB is implemented in C++ and takes advantage of lock-free parallelism and low-level optimizations like memory packing and minimal allocation to achieve its performance. The query engine has several independent query threads (implemented with the pthreads API) which walk the database in parallel. Since the time to query a given PDB structure can vary greatly, a naïve partitioning scheme would not achieve good load balancing. Instead, threads greedily claim the next structure in the database, with lock-free coordination

achieved via atomic operations. Each thread has its own state, so there is no contention during the query process except on the single atomic database cursor.

When searching a model, query threads first search for all candidate atoms as described in Section 3.2. If and only if candidates are found, then a spatial tree is required to search for distance constraints. This tree is constructed in a lightweight in-place fashion by using the coordinate data directly in the database. Thus, only a 'skeleton' of the tree must be created, which minimizes copying and memory allocation.

Upon receiving a query, the web server (itself a C++ program) launches the query threads, with parameters set to indicate the number of results desired and offset for pagination. When the query threads are finished, the web server combines these results into a page and responds to the web browser.

4 Results and discussion

We have developed SpeeDB, a server for searching and analyzing crystal structures of proteins. Queries have been constructed to identify ionic interactions, disulphide bonds, aromatic–aromatic interactions, sulphur–aromatic interactions and cation–π interactions. Matches can be restricted to a specific group of proteins using organism, resolution and/or SCOP classification filters.

4.1 SpeeDB capabilities

4.1.1 Annotated connections

To search for structures that include an annotation connection, users can select the type of connection they are interested in and specify a distance cut-off. The distance cut-off is set to a reasonable default value for each connection type. The connection types available are covalent bond, disulphide bridge, hydrogen bond, metal coordination and ionic interaction. Users can add more than one type of connection to combine the interaction queries, or remove if necessary.

4.1.2 Atom distances

Users can identify protein structures with a specified distance between a pair of atoms in two different amino acids. To add an atom pair to a query, the user selects the amino acids and the two atom types, specifies the chain constraint, and provides the minimum and maximum distance in Å. The distance query can also be branched at different atoms of the first amino acid (root) in order to identify multiple interactions to the same amino acid. In each branch, the atom types of the root amino acid are specified, followed by the amino acids and atom types (leaf atoms) to connect to, and finally the chain and distance constraints for each branch. A single query may contain several atom pairs (or branched atom pairs) along with several annotated connections.

4.1.3 Pre-packaged interaction queries

A query is completely described by a URL. Links for interaction queries described in the previous section are provided as 'Pre-packaged Interaction Queries' in the interface. Following such a link brings the users to the interface with all settings configured appropriately. This way, it is simple to modify existing queries using the SpeeDB interface. For example, it is simple to load the Ionic interaction query, and restrict it to only crystal structures from *Homo sapiens* with a certain resolution.

4.1.4 Filters

Additional filters can be added to restrict matches to certain PDB entries. An organism can be specified by NCBI taxonomy ID (e.g. 9606) or keywords (e.g. 'homo sapiens'). To filter out crystal structures based on quality, the user can specify the resolution cut-off value. The default cut-off value for resolution is 2.5 Å, i.e. only structures with a resolution of less than 2.5Å will be returned. The SCOP classification filter can be used to filter out the protein structures based on the class, fold, super-family or family they belong to. A desired SCOP level can be selected and then a search term entered.

4.1.5 Results

Once the desired query and filters have been specified, the user can submit the query to see the structures that match. The results are shown in a table displaying the PDB ID, name, organism(s), family and resolution of the matching proteins. For each entry, clicking on the PDB ID takes the user to a page displaying more details, clicking on the organism name takes the user to the NCBI taxonomy page, and clicking on the family name takes the user to the SCOP classification database.

4.1.6 Detailed results

The detailed results page shows additional information about a matching protein. A 3D view of the structure is shown along with a link to the corresponding PDB entry. Also shown are the SCOP classification (class, fold, super-family and family), refinement (method of identification, resolution, R-value and R-free) and entities information (entity type, organism and description). Each matching annotated connection and atom pair is shown in a table, with identifying information along with the exact distance found.

4.2 SpeeDB performance

To test the performance and scalability of SpeeDB, queries were run on increasingly large subsets of the database. The resulting timings for the moderately complex Sulphur–Aromatic query are shown in Figure 1. All queries show similar scalability, but absolute time varies with query complexity. Corresponding figures for the other interaction queries are given in Supplementary Figure S3. The speedup for all queries is shown in Figure 2. The machine used has an Intel Core i7-3770 quad-core CPU, with hyper-threading which allows for eight simultaneous hardware supported threads.

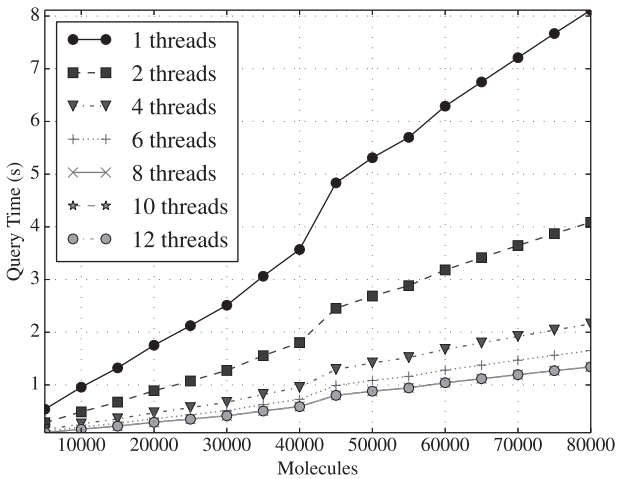

Fig. 1. Query time versus database size for sulphur–aromatic query. Timings show that query time scales roughly linearly with the database size

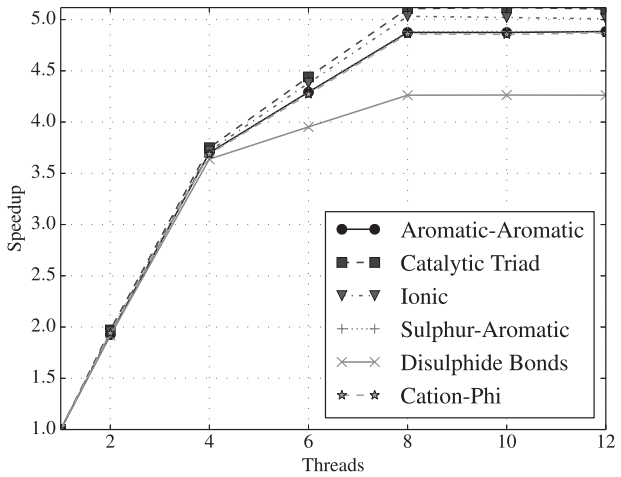

Fig. 2. Speedup versus number of threads for all interaction case studies. This figure is based on the largest benchmark run of 80 000 proteins, e.g. the Sulphur–Aromatic speedup is calculated from the rightmost points in Figure 1. The sharp cut-off at eight threads is due to the fact that SpeeDB is I/O-bound, so there is no advantage to running more threads than the processor supports

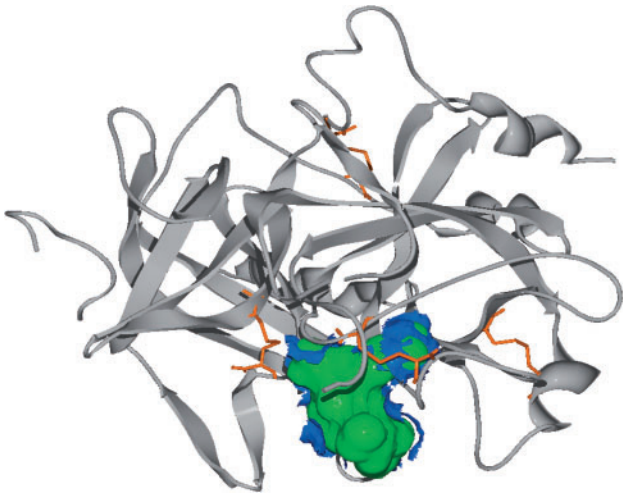

Fig. 3. Analysis of the disulphide bonds in alpha-thrombin (1A3B). The molecular surfaces of the ligands (green) and contact surface (blue) were calculated using Swiss PDB-Viewer. Disulphide bonds are coloured orange to show parts of the secondary structures they connect

These results show that SpeeDB scales predictably as the database size grows, and achieves nearly linear speedup with the use of more cores. Thus, SpeeDB can be expected to adapt well to future improvements in hardware and increases in the PDB database size.

4.3 Case studies

Understanding where protein interactions occur and how they stabilize protein structures helps gain insights into how proteins achieve their native conformation. Correct conformation allows proteins to interact with other biological molecules inside the cellular environment, between cells and also on the cell surfaces. For example, understanding the interactions of amino acids within an active site of an enzyme may help in designing inhibitors or drugs that bind with high affinity to that active site. Studying these interactions on a broader scope (family, super-family or fold) uncovers trends of these interactions in similar or related proteins. This could be useful to identify the most important interactions that play part in a family of proteins.

4.3.1 Distribution of disulphide bonds in human proteases

Proteases are enzymes that break down other proteins and polypeptides by cleaving peptide bonds. They are involved in many intracellular and extracellular processes, including life cycles of pathogens and viruses, where they function in development and progression of diseases (Atkinson *et al.*, 2009; Mittl and Gru, 2006). For these reasons, proteases are considered as good drug targets for many diseases including HIV. An investigation in SpeeDB was carried out to identify the role played by disulphide bonds in human proteases.

To identify high quality structures of human proteases containing disulphide bonds, the disulphide bond query was used with a value of 2.4 Å. 'Homo sapiens' was used as an organism filter to only search structures belonging to humans. The resolution filter was set to 2.5 Å. Structures were filtered to include only those with 'protease' in their SCOP family name. Results with unknown family were hidden and the query was run. A total of 223 protein structure matches were found and 4 structures (1A3B, 1DDJ, 1H1B and 1OWE) were selected at random for further analysis (Fig. 3; Supplementary Fig. S1).

Analysis of the distribution of the disulphide bonds in the selected proteins reveals that the disulphide bonds are located close to the ligand binding sites of proteins. Four disulphide bonds were identified on 1A3B, and three of them are close to the binding site of the ligand (Fig. 3). Similar results were observed with the other three protein structures (Supplementary Fig. S2). Two of the four disulphide bonds identified on 1OWE are also located close to the binding site of the ligand and they seem to participate in the binding of the ligand. Eight disulphide bonds were found on 1HB1 and only four of them are not located close to the binding sites of the ligands. No ligand was associated with 1DDJ, however, looking at the overall structure of the dimeric protein, the disulphide bonds are distributed throughout the protein. The eukaryotic protease family also seems to have disulphide bonds as the predominant type of interaction in this family as shown in Supplementary Fig. S1. The results suggest that the disulphide bonds could play an important role in determining the conformation of the active site which promotes substrate specificity in proteases.

These findings are consistent with the work done by Siddiqui *et al.* (2005), which analyzed the role played by the disulphide bonds on the activity and stability of the cold-adapted α-amylase from *Pseudoalteromonas haloplanktis* bacterium. They used β-mercaptoethanol to reduce the cysteine residues involved in disulphide bond formation or chemically modified them using iodoacetamide or iodoacetic acid. They then analyzed the effects of these modifications on the stability and activity of α-amylase. They identified that there was a loss of stability and activity when the disulphide bonds were broken and blocked. They also found that the effects of the modifications on the enzyme were greater on the enzyme activity as compared to enzyme stability on the active-site region. This demonstrated that the presence of the disulphide bonds in the enzyme is to promote the preservation of the active-site conformation.

4.3.2 Protein interactions in attractive drug target families

The protein kinases and the G protein-coupled receptors (GPCRs) are among the most studied family of proteins, mainly due to the roles they play in regulation of cellular processes. Protein kinases regulate cellular processes through addition of a phosphate group to

a substrate protein. GPCRs on the other hand are responsible for recognition and transduction of intracellular messenger molecules and sensory messages, which allows cells to communicate with each other as well as the environment. Approximately 90% of GPCRs are rhodopsins, which are light-activated and turn on the signalling pathway leading to vision (Palczewski, 2000). The involvement in important cellular processes of proteins from these diverse families has made them attractive drug targets. Studying the interactions that occur within proteins may help us gain insights into their biological properties that may help in drug design.

A search was conducted in SpeeDB to identify the most frequent types of interactions that occur in proteins from the protein kinase and GPCR families. The 'Pre-packaged Interaction Queries' were used for each type of interaction in combination with the organism filter ('Homo sapiens' key word), resolution filter (2.5Å to find high resolution structures) as well as the classification filter ('rhodopsin' and 'kinase' keyword search under family). The results were downloaded as a CSV file and parsed into a Java program that creates a table for the results, which were used in R (http://www.r-project.org/) to analyze the results.

SpeeDB found no human proteins belonging to the GPCR family using the keyword 'rhodopsin'. There were no proteins for human classified under the rhodopsin family with a resolution less than 2.5 Å, so the search to identify the interactions in the GPCR was changed (no organism filter and no resolution filter). The results for the predominant types of interaction in the protein kinase and GPCR families is summarized in Supplementary Figure S1. Twenty-three protein kinase families were identified and most of them contain the ionic and aromatic-based interactions. In the two GPCR (rhodopsin) families identified (not human related), the rhodopsin-like family has disulphide bonds and ionic interactions as the predominant types of interactions, whereas the bacteriorhodopsin-like family contains mostly the aromatic–aromatic and cation–π interactions.

Further analysis of four of the proteins from these families was done to identify the distribution of the interactions and how they are related to the stability and function of the protein. 1F88 was selected from the rhodopsin-like family, 1XIO from the bacteriorhodopsin-like family, 1T45 and 1GKZ were selected from protein kinases catalytic subunit family (Supplementary Tables S1–S4; Fig. S4–S7). The amino acid side chains involved in the interactions were highlighted in the structures. Two disulphide bonds and 37 ionic interactions were identified in the bovine rhodopsin, 1F88 (Supplementary Fig. S4; Table S1). The disulphide bonds are located close to the retinal molecule. According to Palczewski (2000), these disulphide bridges are conserved and contribute towards the stability of the seven-helix transmembrane motif.

A total of 17 aromatic–aromatic, 20 ionic and 2 sulphur–aromatic interactions were found in the *Anabaena* sensory rhodopsin, 1XIO (Supplementary Table S2, Fig. S5). Some of the aromatic residues involved in aromatic–aromatic interactions also seem to interact with the retinal molecule. The analysis by Vogeley *et al.* (2004) revealed that the side chains of Tyr132, Phe139, Trp176, Trp76 and Trp183 are responsible for keeping the retinal molecule in its binding site. Both the bovine and *Anaebena* sensory rhodopsin are orthologous proteins. However, there are differences in the types of interactions found that stabilize both structures. The differences could be responsible for specificity and functional differences between the two rhodopsins.

A total of 12 aromatic–aromatic, 20 cation–π, 62 ionic and 11 sulphur–aromatic interactions were identified in the c-Kit tyrosine kinase, 1T45 (Supplementary Table S3, Fig. S6). The cation–π interaction between Tyr832 and Arg796 found on the activation loop, which is essential to the regulation and activity of protein kinases, in c-Kit functions to keep the enzyme in an autoinhibited conformation (Mol *et al.*, 2004). The cation–π interaction between Tyr832 and Arg796 was also identified in our analysis. Other interactions identified in our analysis that were reported by Mol *et al.* (2004) include the ionic interactions between Arg796 and Asp792 as well as Glu640 and Lys623.

A total of 13 aromatic–aromatic, 17 cation–π, 67 ionic and 6 sulphur–aromatic interactions were identified in the serine/threonine protein kinase B/Atk, 1GZK (Supplementary Table S4, Fig. S7). According to Yang *et al.* (2002), protein kinase B/Atk is activated through phosphorylation at two regulatory sites: the activation loop (Thr309) and the hydrophobic motif (Ser474). However, the activation loop in the structure of protein kinase B/Atk is absent as there was no electron density visible for the segment during structure determination.

The two amino acid residues, Lys181 and Asp293, in protein kinase B/Atk are responsible for coordinating the phosphate groups in the ATP binding site Yang *et al.* (2002). However, in the inactive enzyme, these residues are displaced, causing a disruption of the ATP binding site. The displacement of these residues also seems to cause the two residues to interact, i.e. an ionic interaction was found to be present between Asp293 and Lys181 in our analysis (Supplementary Table S4). The electrostatic interactions observed in protein kinase B/Atk are responsible for peptide associations and form the basis of the protein activation. Both 1T45 and 1GZK represent the inactive catalytic domains of the kinases. The differences in the types of interactions identified in these structure are responsible for the differences in the stability and functions of these kinases.

5 Conclusion

SpeeDB is an online resource that allows users to quickly identify proteins structures using different types of interaction queries. SpeeDB includes queries to identify various interactions, including disulphide bonds, aromatic–aromatic, sulphur–aromatic, cation–π and ionic interactions. Additionally, users can construct custom queries based on atom characteristics, distances between atoms on different amino acids, and 'annotated connections' described in the protein structure files. Organism, resolution and classification filters can be added to queries to filter out proteins of interest.

Results can be shown as a web page, or downloaded in CSV format. The ability to process results in other tools via CSV is particularly useful when analysing large result sets. As shown in Section 4.3.1, the distribution of PDB matches to a query can be very informative. However, care must be taken in interpreting the distribution. First, although PDB is large, it is not statistically representative since entries have been added to PDB ad hoc as various researchers resolved structures. Moreover, SpeeDB searches all available structures and does not attempt to filter duplicate and near-duplicate structures from the results. In the future, SpeeDB may be enhanced to deal directly with these issues. Its current aim is to quickly find structures of interest and, in conjunction with possible post-processing of the results, SpeeDB is useful for answering new biological questions as demonstrated by the case studies presented here.

Acknowledgements

The authors thank the members of the Protein Structure Function Research Unit (PSFRU), Pierre Durand and Nichole Cerutti, all from the University of

Witwatersrand, for their support and constructive feedback through the development of this resource.

Funding

S.H. is partially supported by a National Research Foundation research incentive award. F.D. and D.R. are partially supported by the Natural Sciences and Engineering Research Council of Canada and the IBM Center for Advanced Studies Canada.

Conflict of Interest: none declared.

References

Atkinson,H. J. *et al.* (2009) The global cysteine peptidase landscape in parasites. *Trends Parasitol.*, **25**, 573–581.

Babu,M.M. (2003) NCI: a server to identify non-canonical interactions in protein structures. *Nucleic Acids Res.*, **31**, 3345–3348.

Barlow,D. J. and Thornton,J. M. (1983) Ion-pairs in proteins. *J. Mol. Biol.*, **168**, 867–885.

Bentley, J.L. (1975) Multidimensional binary search trees used for associative searching. *Commun. ACM*, **18**, 509–517.

Berman, H.M. *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Burley,S. and Petsko,G. (1986) Amino-aromatic interactions in proteins. *FEBS Lett.*, **203**, 139–143.

Burley,S.K. and Petsko,G.A. (1985) Aromatic-aromatic interaction: a mechanism of protein structure stabilization. *Science (New York, N.Y.)*, **229**, 23–28.

Cerutti,N. *et al.* (2010) Stabilization of HIV-1 gp120-CD4 receptor complex through targeted interchain disulfide exchange. *J. Biol. Chem.*, **285**, 25743–25752.

Chourasia,M. *et al.* (2011) Aromatic-aromatic interactions database, A(2)ID: an analysis of aromatic π-networks in proteins. *Int. J. Biol. Macromol.*, **48**, 540–552.

Fersht,A.R. and Serrano,L. (1993) Principles of protein stability derived from protein engineering experiments. *Curr. Opin. Struct. Biol.*, **3**, 75–83.

Goldenberg,D.P. (1985) Dissecting the roles of individual interactions in protein stability: lessons from a circularized protein. *J. Cell. Biochem.*, **29**, 321–335.

Hazelhurst,S. (2010) PH2: An Hadoop-based framework for mining structural properties from the PDB Database. In: *Proceedings of the 2010 Annual Research Conference of the South African Institute of Computer Scientists and Information Technologists*, SAICSIT '10, ACM, New York, NY, pp. 104–112.

Manikandan,K. and Ramakumar,S. (2004) The occurrence of C–H ... O hydrogen bonds in alpha-helices and helix termini in globular proteins. *Proteins*, **56**, 768–781.

Meyer,E. *et al.* (2003) Interactions with aromatic rings in chemical and biological recognition. *Angewandte Chemie*, **42**, 1210–1250.

Mihel,J. *et al.* (2008) PSAIA - protein structure and interaction analyzer. *BMC Struct. Biol.*, **8**, 21.

Mittl,P.R.E. and Gru,M.G. (2006) Opportunities for structure-based design of protease-directed drugs. *Curr. Opin. Struct. Biol.*, **16**, 769–775.

Mol,C.D. *et al.* (2004) Structural basis for the autoinhibition and STI-571 inhibition of c-Kit tyrosine kinase. J. Biol. Chem., **279**, 31655–31663.

Pal,A. *et al.* (2009) IntGeom: a server for the calculation of the interaction geometry between planar groups in proteins. *J. Proteom. Bioinform.*, **02**, 60–63.

Palczewski,K. (2000) Crystal Structure of Rhodopsin: A G Protein-Coupled Receptor. *Science*, **289**(5480), 739–745.

Pitre,S. *et al.* (2006) PIPE: a protein-protein interaction prediction engine based on the re-occurring short polypeptide sequences between known interacting protein pairs. *BMC Bioinformatics*, **7**, 365.

Reid,K.S.C. *et al.* (1985) Sulphur-aromatic interactions in proteins. *FEBS Lett.*, **190**, 209–213.

Sathyapriya,R. and Vishveshwara,S. (2004) Interaction of DNA with clusters of amino acids in proteins. *Nucleic Acids Res.*, **32**, 4109–4118.

Scrutton,N. and Raine,A. (1996) Cation-pi bonding and amino-aromatic interactions in the biomolecular recognition of substituted ammonium ligands. Biochem. J., **319**, 1–8.

Siddiqui,K. *et al.* (2005) Role of disulfide bridges in the activity and stability of a cold-active α-amylase. *J. Bacteriol.*, **187**, 6206–6212.

Sowdhamini,R. *et al.* (1989) Steriochemical modelling of disulfide bridges. Criteria for introduction into proteins by site directed mutagenesis. *Protein Eng.*, **3**, 95–103.

Stickle,D.F. *et al.* (1992) Hydrogen bonding in globular proteins. *J. Mol. Biol.*, **226**, 1143–1159.

Sundaramurthy,P. *et al.* (2010) HORI: a web server to compute higher order residue interactions in protein structures. *BMC Bioinformatics*, **11** (Suppl. 1), S24.

Tina,K. *et al.* (2007) PIC: protein interactions calculator. *Nucleic Acids Res.*, **35**(Web Server issue), W473–W476.

Vogeley,L. *et al.* (2004) Anabaena sensory rhodopsin: a photochromic color sensor at 2.0 A. *Science (New York, N.Y.)*, **306**, 1390–1393.

Westbrook,J.D. and Bourne,P.E. (2000) STAR/mmCIF: An ontology for macromolecular structure. *Bioinformatics*, **16**, 159–168.

Yang,J. *et al.* (2002) Molecular mechanism for the regulation of protein kinase B/Akt by hydrophobic motif phosphorylation. *Mol. Cell*, **9**, 1227–1240.