BIOINFORMATICS

Vol. 20 no. 16 2004, pages 2767–2777
doi:10.1093/bioinformatics/bth326

ICBS: a database of interactions between protein chains mediated by β-sheet formation

Yimeng Dou[1,2,†], Pierre-François Baisnée[1,2,†], Gianluca Pollastri[1,2,‡], Yann Pécout[1,2,5], James Nowick[2,3] and Pierre Baldi[1,2,4,*]

[1]School of Information and Computer Science, [2]Institute for Genomics and Bioinformatics, [3]Department of Chemistry and [4]Department of Biological Chemistry, University of California, Irvine, CA 92697, USA and [5]IUP Génie Physiologique et Informatique, University of Poitiers, 86000 Poitiers, France

Received on January 1, 2004; revised on March 4, 2004; accepted on May 5, 2004
Advance Access publication May 27, 2004

ABSTRACT

Motivation: Interchain β-sheet (ICBS) interactions occur widely in protein quaternary structures, interactions between proteins and protein aggregation. These interactions play a central role in many biological processes and in diseases ranging from AIDS and cancer to anthrax and Alzheimer's.

Results: We have created a comprehensive database of ICBS interactions that is updated on a weekly basis and allows entries to be sorted and searched by relevance and other criteria through a simple Web interface. We derive a simple ICBS index to quantify the relative contributions of the β-ladders in the overall interchain interaction and compute first- and second-order statistics regarding amino acid composition and pairing at different relative positions in the β-strands. Analysis of the database reveals a 15.8% prevalence of significant ICBS interactions, the majority of which involve the formation of antiparallel β-sheets and many of which involve the formation of dimers and oligomers. The frequencies of amino acids in ICBS interfaces are similar to those in intrachain β-sheet interfaces. A full range of non-covalent interactions between side chains complement the hydrogen-bonding interactions between the main chains. Polar amino acids pair preferentially with polar amino acids and non-polar amino acids pair preferentially with non-polar amino acids among antiparallel (i, j) pairs. We anticipate that the statistics and insights gained from the database will guide the development of agents that control interchain β-sheet interactions and that the database will help identify new protein interactions and targets for these agents.

Availability: The database is available at: http://www.igb.uci.edu/servers/icbs/

Contact: pfbaldi@ics.uci.edu

Supplementary Information: http://www.igb.uci.edu/servers/icbs/

1 INTRODUCTION

Interactions between the hydrogen-bonding edges of β-strands in different protein chains constitute an important mode of protein–protein interaction. In this under-appreciated mode of molecular recognition between proteins, hydrogen bonds form between β-strands belonging to different protein chains. The partnering β-strands are studded with an alternating array of hydrogen-bond-donors and acceptors with the pattern donor-acceptor-space, etc. These interchain β-sheet (ICBS) interactions, which are represented schematically in Figure 1A, are frequent in protein dimers and other quaternary structures, in interactions among different proteins, and in protein aggregation.

In addition to occurring in non-pathological protein interactions, ICBS interactions are involved in diseases, ranging from cancer and AIDS to anthrax and Alzheimer's (Maitra and Nowick, 2000): Ras oncoproteins activate Raf kinase enzymes in part through edge-to-edge interactions between β-sheet structures (Nassar *et al*., 1995), as shown in Figure 1B. HIV-1 protease functions as a dimer in which two 99mer monomers self-assemble, in part through the interdigitation of the N- and C-terminal β-strands to form a four-stranded antiparallel β-sheet (Wlodawer *et al*., 1989). The anthrax toxin protective antigen delivers the anthrax lethal and edema factors to the host cell by self-assembling in the cell membrane through ICBS interactions to generate a heptamer which binds these factors (Petosa *et al*., 1997). ICBS interactions occur in the protein aggregation that is involved in generating β-amyloid in Alzheimer's disease and in many other neurodegenerative diseases, including Huntington's disease, Creutzfeld-Jacob disease, and other prion diseases. ICBS interactions occur commonly in domain swapping and related strand complementation processes, such as the formation of PapD–PapK

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

‡Present address: Department of Computer Science, University College Dublin, Ireland.

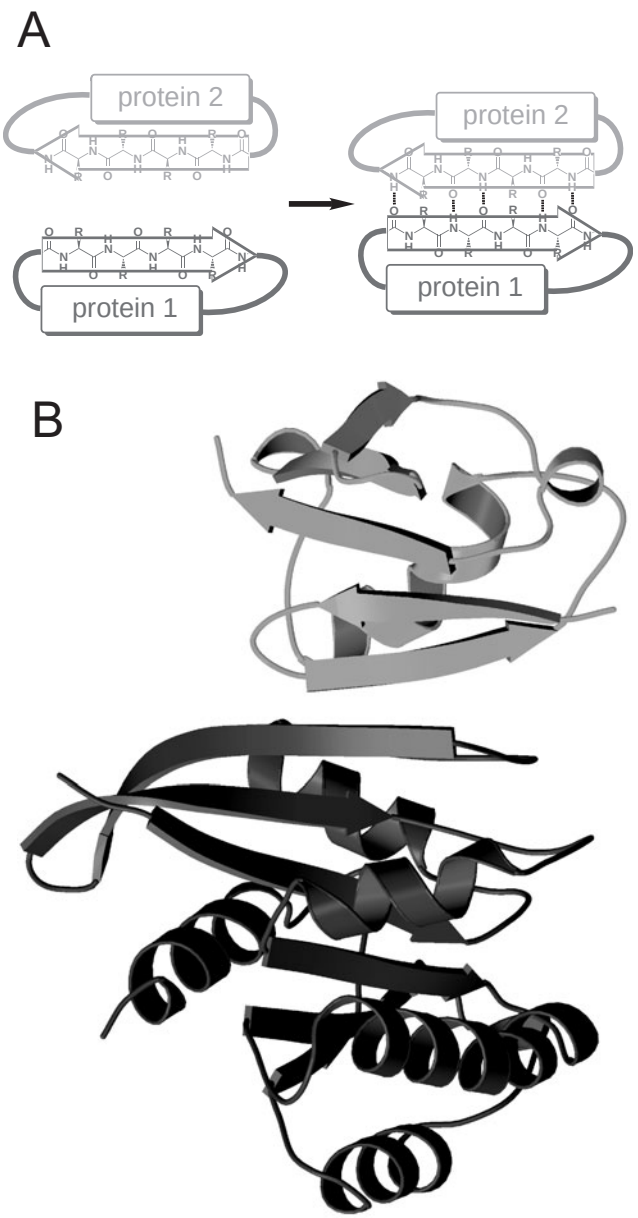

Fig. 1. (**A**) Schematic Representation of Interchain β-Sheet Interactions. (**B**) Representative ICBS Interaction. Complex between the Ras-related protein Rap1A and the serine/threonine kinase c-Raf1 (PDB ID 1gua). Diagram was prepared with MolScript (Kraulis, 1991) and Raster3D (Merritt and Bacon, 1997).

and FimC–FimH chaperone–adhesin complexes associated with uropathogenic *Escherichia coli* (Sauer *et al*., 1999; Choudhury *et al*., 1999) and in the binding of PDZ domains to membrane receptor and ion channel proteins (Doyle *et al*., 1996).

Subtle changes in ICBS-based molecular recognition can often have profound consequences for health and disease. Gain-of-function mutations in platelet receptor glycoprotein Ibα (GpIbα) that strengthen the β-sheet interactions with the plasma protein von Willebrand factor (VWF) give rise to congenital bleeding disease (von Willebrand disease type 2B) (Huizinga *et al*., 2002). Several of these point mutations occur in the loop comprising residues 227–241 of GpIbα and increase its propensity to adopt a β-hairpin conformation and form an interchain β-sheet with the β3 strand of VWF domain A1 (residues 562–566). These mutations increase the affinity of the two proteins for each other and result in disease. In transthyretin (TTR) amyloid diseases, such as familial amyloid polyneuropathy, mutations that destabilize the TTR β-sheet tetramer and facilitate its dissociation (e.g. V30M) give rise to disease (Hammarström *et al*., 2001). A second mutation at the β-sheet interface between TTR monomers that stabilizes the tetramer and impedes dissociation (T119M) protects heterozygous individuals with both mutations from developing the disease.

In spite of the importance of ICBS interactions among proteins, there have been no systematic methods of identifying them. At the most straightforward level, 'identification' could consist of sifting through known protein structures in the Protein Data Bank (PDB, http://www.rcsb.org/pdb/) (Berman *et al*., 2000; Westbrook *et al*., 2002) and flagging and characterizing those structures with ICBS interactions. Such a process would, for example, identify published structures such as those of the HIV-1 protease dimer (e.g. PDB identifier 9hvp) (Wlodawer *et al*., 1989) and the complex between the Ras-related protein Rap1A and the serine/threonine kinase c-Raf1 (PDB identifier 1gua, Figure 1B) (Nassar *et al*., 1995). Identification of ICBS interactions among proteins lacking high-resolution structures (e.g. β-amyloid), among proteins in which the conditions required for ICBS interaction are different than those used for collection of structural data (e.g. the anthrax toxin protective antigen), or among proteins for which structures of the individual partners are only present separately in the PDB, would constitute a considerably more difficult challenge.

The present work overcomes the simpler problems associated with identifying ICBS interactions by creating a comprehensive relational database of ICBS interactions that is updated on a weekly basis, can identify quaternary structures, and allows the entries to be ranked and searched by relevance and other criteria through a simple Web interface available through www.igb.uci.edu/servers/icbs/. The ICBS database identifies and characterizes all interchain β-sheet interactions within entries in the PDB and the Protein Quaternary Structure server (PQS, http://pqs.ebi.ac.uk/) (Henrick and Thornton, 1998), of the Macromolecular Structure Database of the European Bioinformatics Institute (EBI).

2 METHODS

This section describes the development of the ICBS database: the choice of protein structure sources, the identification and characterization of ICBS interactions, and the creation of the user interface.

2.1 Data sources

Protein structures were obtained from the PDB and the PQS server. The PDB stores all experimentally determined protein structures. However, a PDB structure does not always correspond to the known or likely biological form of a protein, which may contain multiple copies of the asymmetric unit the PDB structure describes. Symmetry operations must then be applied to the asymmetric unit to obtain the quaternary structure. This is for instance the case of a modified HIV-1 protease (3hvp), whereas another PDB entry for a related HIV-1 protease complex (9hvp) contains the full macromolecule. The PQS server automatically generates likely quaternary structures for PDB protein entries. To spot interchain β-sheet (ICBS) interfaces within known structures, we therefore use both the experimentally determined PDB coordinates, and the coordinates of the corresponding likely quaternary structures. The former are obtained from the PDB ftp site (ftp://ftp.rcsb.org/pub/pdb/data/structures/), the latter from the PQS http site (http://pqs.ebi.ac.uk/).

PDB structures are available in both the original PDB file format—still representing the official release of PDB structures on the ftp site—and in a more recent mmCIF file format—available only from the PDB beta ftp site at the time of this writing. Not only do the mmCIF files overcome format limitations that PDB files suffer from, they also reflect the current, revised content of the PDB database, after the results of the Data Uniformity Project (Westbrook *et al*., 2002). However, PQS structures are currently available as PDB files only, and mmCIF files do not yet represent an established source for PDB structures. For the initial version of the ICBS database, we therefore chose to use PDB files, despite the disadvantages resulting from their fixed-field-width format and the errors they contain. When both PQS and PDB structures become officially available as mm-CIF files or through direct database queries, a second version of the ICBS database will be released, which will solely rely on the improved formats and data sets.

2.2 Identification of ICBS interactions

Through automated procedures implemented in PERL scripts, we update the ICBS database weekly. We first download and scan every newly deposited or revised PDB entry and the (possibly multiple) corresponding likely quaternary structures. We consider only the PQS-entry types that are relevant to the ICBS database, namely the 'SYMMETRY-COMPLEX' and 'SPLIT-SYMMETRY' types.

For each multimeric candidate structure, we derive secondary structure elements from the three-dimensional (3D) structure, using the 'Define Secondary Structure of Proteins' (DSSP) program (Kabsch and Sander, 1983) in its updated version of April 1, 2000, available from the Centre for Molecular and Biomolecular Informatics (CMBI).

In the output that the DSSP program generates, we then spot β-ladders (a set of one or more consecutive β-bridges of the same type) that join residues belonging to different chains. We thus compile a list of proteins with interchain β-ladder interactions, along with a list of pairing chains. To characterize the interactions, we compute an ICBS index and compile information from the PDB and DSSP data, as explained below. It is worth noting that we consider single β-bridges, i.e. ladders of length 1. Single β-bridges can strengthen the effect of regular ladders (of length > 1), and they are therefore taken into account for the computation of the ICBS index. While we include in the database structures whose interchain interfaces display only one or more single, isolated β-bridges, we offer an option to exclude such structures when querying the database.

New or modified ICBS entries are finally stored in an open-source relational database—implemented using mySQL (http://www.mysql.com/)—while discarding any obsolete or superseded structure. The weekly updates are scheduled at such a time that the new PQS structures corresponding to new or modified PDB entries should be present on the PQS server. However, to compensate for possible delayed releases of PQS structures, the temporal scope of each weekly update extends several weeks in the past.

2.3 The ICBS index

To characterize and rank ICBS interactions, we measure the relative importance of the β-ladders in the overall interchain interface between two chains A and B through the following simple ICBS index:

$$I_{AB} = 1000\frac{H_{AB}}{C_{AB}} \quad (1)$$

where H_{AB} represents the number of hydrogen bonds in the β-bridges that join A and B, and where C_{AB} represents the total number of heavy atom (non-hydrogen atom) contacts between A and B. Here, the number of hydrogen bonds is estimated as the number of DSSP β-bridges contributing to the interchain interface, over all ladders, inclusive of single-bridge ladders. The number of heavy atom contacts is derived from the 3D coordinates by computing pairwise atom distances between the chains A and B. Two heavy atoms belonging to different chains are here considered as being in contact when they are $<$4.4 Å apart. This corresponds to twice the length of a C–H bond (1.1 Å), plus twice the van der Waals radius of H (1.1 Å), i.e. to the maximum expected distance between the carbons of two CH groups; for O and N atoms, the distances should be shorter. Contacts involving hydrogen atoms are not counted, to preserve the comparability of indices computed on crystallographic and on nuclear magnetic resonance (NMR) structures. The higher the ICBS index for a chain pair (A, B), the higher the structural importance of the β-sheet(s) in the overall interactions between the two chains.

To characterize the strength of the interchain β-sheet interactions in a protein structure S, we retain the maximum index

I_{AB} over all chain pairs in the protein:

$$I_S = \max_{A,B} I_{AB} \quad (2)$$

Although we experimented with various thresholds for the definition of an atom contact, and tried $C_\alpha - C_\alpha$ and $C_\beta - C_\beta$ contacts instead of heavy atom contacts as a fast approximation, we found the index based on hydrogen bonds and heavy atom contacts to be the most informative and reliable.

2.4 Homogeneity and orientation of β-sheet interfaces

Two additional characteristics of the ICBS interfaces are derived from the analysis of the DSSP output: their homogeneity and their orientation. The interface between two chains is homogeneous when the chains are identical and is heterogeneous when the chains are not. As a computational shortcut for the first version of the database, we assume that two chains are identical when they have the same number of residues; homology and stereochemistry criteria will be added in the next version, with very few changes expected. At the level of the whole protein, the interface is either homogeneous or heterogeneous when all chain-pair interfaces fall in the same homogeneity category, and it is mixed when β-ladders exist between both identical and different chains.

Likewise, the orientation of the interface between two chains can be parallel, antiparallel or mixed, depending on the type of the ladders that join the chains. The same categories apply at the protein level for a global characterization of the orientation of the ICBS interfaces the protein contains.

2.5 Additional characterization

To let users filter or sort ICBS entries in multiple ways, we also store in the database additional information, such as the number of hydrogen bonds and heavy atom contacts between any pair of chains, the number of chains and residues in the protein, the experimental technique to determine the structure, the journal that published the primary citation for the structure, and the PDB header and code, as well as the deposition and revision dates of the PDB structure corresponding to the entry.

2.6 User interface

The Web interface to the database is implemented using PHP, which is a widely used general-purpose, open-source scripting language that is especially suited for Web development (http://www.php.net). A query form lets users select or exclude entries based on criteria set on any of the pieces of information stored in the database. In particular, it is possible to select or exclude PQS structures whose ICBS interfaces are identical to those already present in the corresponding PDB structures, and which are in this respect redundant. ICBS interfaces consisting solely of single β-bridges can also be selected or excluded (see also Appendix). The results of a query are displayed in one or more html pages. Buttons are provided for sequential or random access to any results page. Results can be sorted according to any relevant piece of information, such as the ICBS index or the homogeneity and orientation of the ICBS interface. The results are linked to the corresponding PDB coordinates and summary information files and to the PQS files where appropriate.

3 RESULTS

Below, we first study the prevalence of ICBS interactions within known structures, and the ranking/distribution of ICBS interactions through the ICBS index. We then show how the ICBS database can be mined and queried effectively, for instance to discover new and potentially interesting ICBS interactions or to analyze how other non-covalent forces can complement ICBS hydrogen bonding in stabilizing protein interactions. Finally, we look at the statistical properties of ICBS and discuss issues of molecular design.

3.1 Prevalence of ICBS interactions

Interchain β-sheet interfaces are widespread, and more so than was initially foreseen by the authors. Of 21 998 protein, peptide, and related structures in the PDB (as of August 11, 2003), 3735 (17.0%) had ICBS interfaces. An additional 1966 structures lacked ICBS interfaces but had one or more PQS entries with such interfaces, making the total fraction of protein structures with ICBS interfaces 25.9%. A significant fraction of these structures have only limited interchain β-sheet contact, consisting only of isolated β-bridges, as determined by DSSP analysis. If this fraction is excluded, these numbers drop to 10.7 and 15.8%, the latter of which probably best describes the prevalence of interchain β-sheet interfaces in protein interactions. The majority of these interfaces are antiparallel. Of the subset of PDB and non-redundant PQS structures that excludes those structures with only isolated β-bridges, 67.9% are exclusively antiparallel, 9.0% are exclusively parallel, and 23.1% contain both antiparallel and parallel components. Some of these structures have more than two pairing chains. If the interactions between all of the chain pairs are considered separately, 75.8% are exclusively antiparallel, 16.6% are exclusively parallel, and 7.6% contain both antiparallel and parallel components. Likewise, 56.6% are homogeneous, i.e. involve identical chains (dimers and oligomers), and 43.4% are heterogeneous.

3.2 Ranking of ICBS interactions

The ICBS index provides a rough, but effective quantification of the relative contribution of the interchain β-sheet contacts to the overall interchain contacts. For example, the contact between the monomeric units of the defensin HNP-3 dimer (Hill *et al.*, 1991) primarily involves contacts between the interacting β-sheets and gives an ICBS index value of 44.0, while the met repressor dimer (Rafferty *et al.*, 1989) involves contacts between helices as well as those between the interacting β-strands and gives an ICBS index value of 21.7.

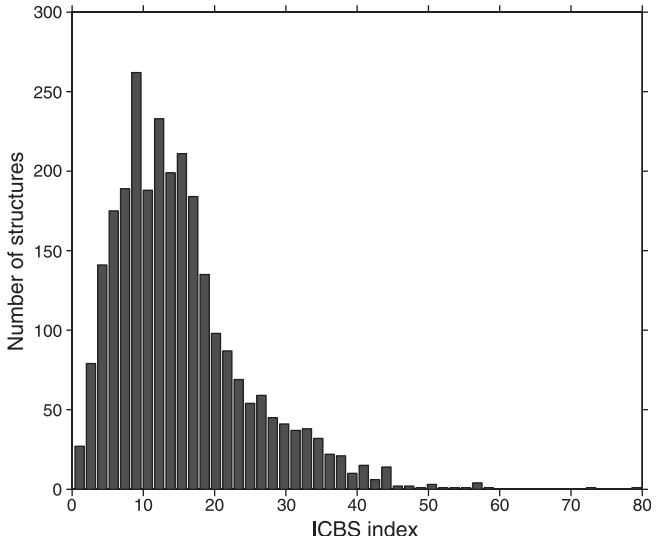

Fig. 2. Distribution of the ICBS Index. Structures whose ICBS interfaces consist solely of single β-bridges and redundant PQS entries are excluded. The histogram represents 2689 structures.

The average ICBS index value, over the subset of PDB and non-redundant PQS structures with interchain β-ladders more than one β-bridge long, is 15.4, with values ranging from 0.3 to 80.0. The histogram of Figure 2 illustrates the index distribution over this subset of the ICBS database. Proteins with interesting interchain β-sheet contacts were found to generally have ICBS index values of ∼10–60.

3.3 A sample database query

The following search illustrates how the ICBS database may be used. In this search, we set out to identify important new ICBS protein interactions that we had not previously identified. We searched for references to interactions between different protein chains that had been published in the leading journals such as *Cell*, *Nature*, *Proceedings of the National Academy of Sciences* and *Science*. To keep the hits simple and easy to interpret, we limited the number of chains to two and the number of residues to 300. A total of 28 hits were obtained through the ICBS database on August 11, 2003, with ICBS Index values ranging from 42.0 to 5.8. Further exploration of these hits revealed, for instance, that interchain β-sheets are important in the interaction of a chromatin-associated protein with a histone (Jacobs and Khorasanizadeh, 2002) and support the 'histone-code' hypothesis, in which modification of histones systematically results in control of gene expression and heritable genetic changes (Pedersen *et al*., 1998; Jenuwein and Allis, 2001). This important role of ICBS interactions represented new knowledge for the authors. It is anticipated that other users will be able to generate new knowledge and interrelationships of ideas about ICBS interactions in a similar fashion.

3.4 Non-covalent forces in ICBS interactions

β-Sheet interactions between proteins involve hydrogen bonding between the main-chain amide groups and many non-covalent contacts between the side chains, including salt bridges, hydrogen bonds, hydrophobic interactions, and van der Waals interactions. These interactions are nicely illustrated by many of the small dimers in the ICBS database, such as interleukin 8 (IL-8) (Clore *et al*., 1990; Baldwin *et al*., 1991), the met repressor (Rafferty *et al*., 1989), defensin HNP-3 (Hill *et al*., 1991), and barstar (Ratnaparkhi *et al*., 1998).

The interleukin 8 dimer, which is representative of the mode of dimerization of the family of CXC chemokines, provides a good introduction to these interactions (Clore *et al*., 1990; Baldwin *et al*., 1991). Interleukin-8 is a chemokine that is released by monocytes, fibroblasts, endothelial cells and keratinocytes upon inflammation to attract neutrophils and T-cells. Each half of the dimer comprises a three-stranded antiparallel β-sheet and a helix (Fig. 3A). The edges of the β-sheets come together in an antiparallel fashion at the dimerization interface, forming six hydrogen bonds between residues 23–29 of the main chains of the monomers (Fig. 3B). Salt bridges between the two Glu24-Arg26 pairs are prominent on the 'top' face of the dimer (Fig. 3C), while a hydrophobic cluster involving both sets of Leu25 and Val27 is noteworthy on the 'bottom' face of the dimer (Fig. 3D). More subtle hydrophobic packing of Ile28 and the hydrophobic regions of Glu24 and Arg26 also contribute to the structure of the dimer. Packing interactions between these interfacial residues and the residues in the adjacent strand further buttress the β-sheet interface. That the dissociation constant (K_d) of the IL-8 dimer is relatively high (18×10^{-6} M), and that dimerization is not necessary for biological activity, provide a reminder that the observation of ICBS interactions is not necessarily proof of their biological significance (Burrows *et al*., 1994; Rajarathnam *et al*., 1994).

The met repressor, which functions as a β-sheet dimer, illustrates the importance of hydrogen bonding among side chains in ICBS interactions (Rafferty *et al*., 1989). The met repressor regulates the biosynthesis of methionine, repressing genes coding for methionine and S-adenosylmethionine (SAM) biosynthesis by binding to the major groove of DNA. The met repressor also binds SAM, which serves as a co-repressor and is thus also involved in regulation of methionine biosynthesis. A large β-sheet interface, comprising 10 hydrogen bonds among residues 20–30, is buried in the major groove upon binding. Among the many ICBS contacts at the interface is a hydrogen-bonded pair of threonine residues (Thr25), which is present in the met repressor dimer when it is not bound to SAM. Upon binding SAM, the distance between the threonine residues increases. When the met repressor-SAM corepressor complex binds to DNA the distance between the threonine residues further increases and the threonine residues hydrogen bond to adenine bases of the DNA. These observations

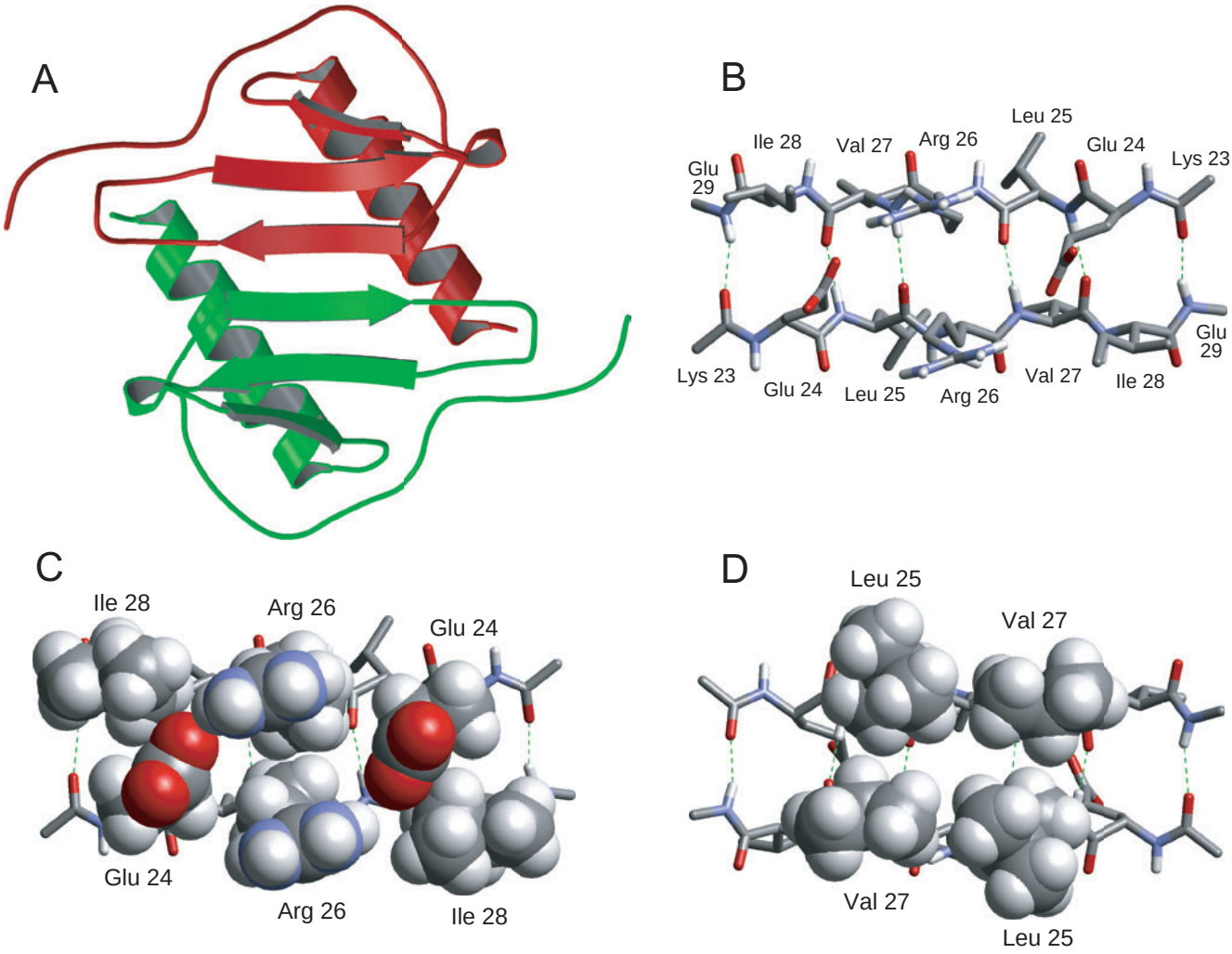

Fig. 3. The Interleukin-8 Dimer (PDB ID 1il8). (**A**) Structure of the dimer. (**B**) View of the dimerization interface illustrating hydrogen bonding between the main chains of the interacting β-sheets. (**C**) View of the dimerization interface illustrating contacts between the side chains on the 'top' face. (**D**) View of the dimerization interface illustrating contacts between the side chains on the 'bottom' face.

suggest that the Thr–Thr hydrogen bond plays a key role in the regulation process, helping to transduce the SAM-binding signal into DNA binding and gene regulation.

ICBS interactions often involve contacts between remote residues, as well as those in adjacent chains. The hydrophobic surfaces of the two halves of the defensin HNP-3 dimer fit snugly together in a 'Yin-Yang' fashion and feature Cys disulfide linkages and Trp, Phe, and Tyr residues (Hill *et al*., 1991). The dimer is thought to be the biologically active form of this microbicidal protein that is produced by human neutrophils to destroy pathogens. Very different sorts of interactions occur in the dimer of the bacterial ribonuclease inhibitor barstar, which exhibits several salt bridges between Glu, Lys and Arg residues that are remote from the ICBS interface (Ratnaparkhi *et al*., 1998). With a relatively small dimerization interface, consisting primarily of the salt bridges, four hydrogen bonds at the ICBS interface, and contacts between the side chains of the residues at the ICBS interface, the barstar dimer exhibits a dissociation constant $K_d = 3 \times 10^{-3}$ M (Korchuganov *et al*., 2001).

3.5 Statistical properties of ICBS and molecular design

We also compute a number of statistics on ICBS interactions and include them in the ICBS database and Web interface. These comprise first-order statistics, i.e. the frequencies of occurrence of each residue in β-strands involved in ICBS interactions, and second-order statistics of amino acid pairings (see also Hutchinson *et al*., 1998) such as the frequencies of each possible pair of amino acids considering positions (i, j), $(i, j+1)$,$(i, j+2)$,$(i, j-1)$,$(i, j-2)$, the ordering on

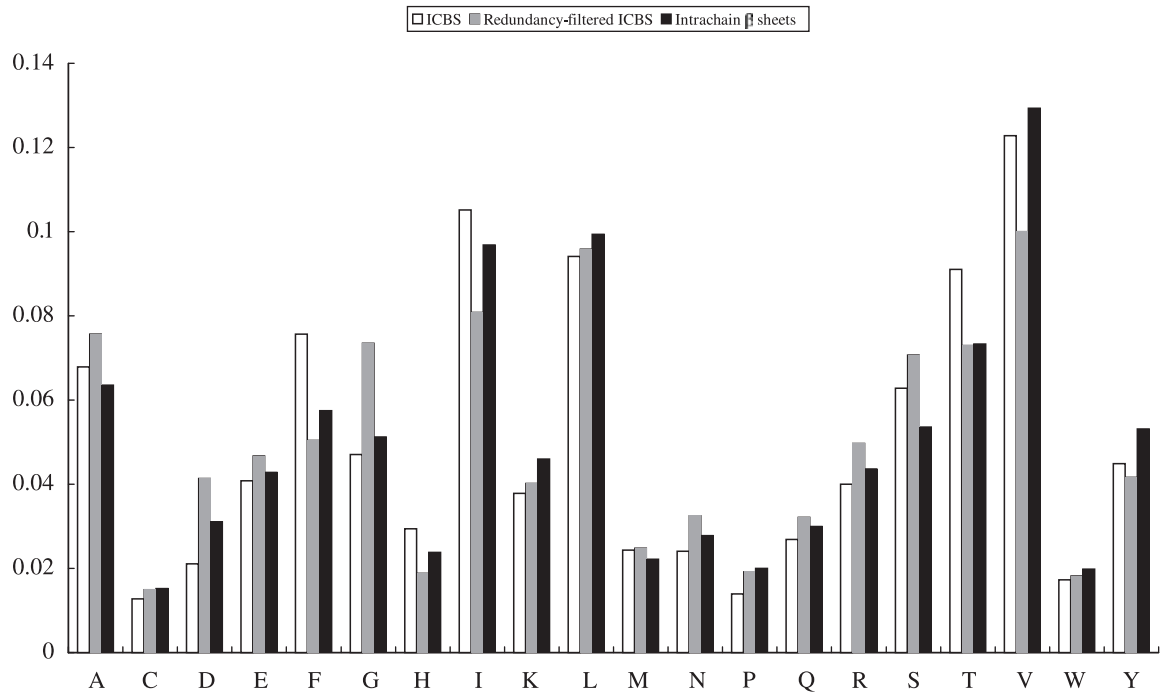

Fig. 4. Frequencies of Amino Acids in Inter- and Intrachain β-Sheets. For each amino acid, the first (white) bar represents the frequency among the entire set of interchain β-sheets, the second (grey) bar represents the frequency among the redundancy-filtered set of interchain β-sheets, and the third (black) bar represents the frequency among intrachain β-sheets.

each strand being from the N- to the C-terminus. These statistics can be restricted to particular subsets of interactions (e.g. parallel versus antiparallel) and compared to similar statistics computed on other data sets, such as the average composition of the Swiss-Prot database or of β-strands involved in intrachain interactions. In addition to raw values and probabilities, for significance analysis in the case of second-order statistics we also report, and allow sorting by, the value of $P(AB)/P(A)P(B)$ computed for each pair of amino acids.

There are numerous biases in the PDB—for instance, it contains well over 100 variants of HIV protease and streptavidin. Thus, in addition to raw statistics, we also compute statistics on a redundancy-filtered set. To filter the redundancy, we consider each pair of ICBS-interacting strands, and add a flank of five amino acids on each side to each strand. We then perform a global alignment of each interacting pair to all the other interacting pairs with a threshold similarity set at 30% identical residues. To compute non-redundant statistics, each interacting pair in a similarity class is weighted by the inverse of the numbers of elements in the class. Thus, in a class containing only two similar ICBS-interacting pairs (X, Y) and (X', Y'), the statistics extracted from the strands X and Y will be weighted by 0.5 and the same holds for the statistics extracted from the strands X' and Y'. Weighting similar objects should retain slightly more information than a more radical approach that would discard entirely redundant similar copies (we use the latter approach only for chains that are identical).

In the current redundancy-filtered set, there are 17 962 antiparallel ICBS and 6596 parallel ICBS with weighting factors ranging from 1 to 532. A complete analysis of these statistics is beyond the scope and length of this paper and here we point out only a few salient observations. To a first-order approximation, the amino acid frequencies in the redundant and non-redundant set are quite similar. They are also similar to the statistics computed on intrachain β-sheets in the PDB (Fig. 4). Similar results are obtained when the frequencies are computed on parallel or antiparallel strands only.

The redundancy-filtered data are essential for meaningful second-order statistics. The pair QN, for example, is highly over-represented in the raw ICBS (i, j) data for antiparallel strands. The high frequency of this pairing results in part because QN occurs in most or all of the HIV-1 protease structures and there are two QN pairings in the HIV-1 protease dimer. This overresentation is addressed by the redundancy filtering operation which reduces the $P(QN)/P(Q)P(N)$ probability score for antiparallel strands approximately from 4.64 to 1.15.

Second-order statistics are particularly relevant for a better understanding of ICBS interactions and may prove valuable for the design of ligands that bind proteins through ICBS interactions. Among antiparallel (i, j) pairs in the

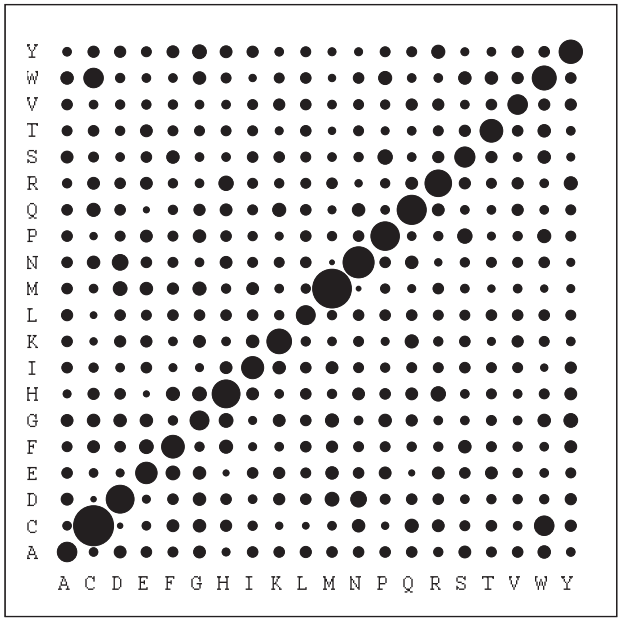

Fig. 5. Relative frequencies of i, j pairings of amino acids among the redundancy-filtered set of antiparallel interchain β-Sheets. The area of each circle represents the relative magnitude of the $P(AB)/P(A)P(B)$ value for each i, j pairing.

redundancy-filtered data, homopairs (e.g. Ala–Ala or Gly–Gly) are generally preferred strongly over heteropairs (e.g. Ala–Gly) (Fig. 5). This preference reflects the large number of homodimers that are present in the database, and is largely absent in antiparallel intrachain β-sheets and in parallel interchain β-sheets. A bias towards self-pairing is also observed to some extent for the $(i, j+2)$ or $(i, j-2)$ positions in antiparallel strands, for similar reasons. The preponderance of homopairs is not eliminated by redundancy filtering, because many of the homodimers are non-redundant.

Of the antiparallel (i, j) homopairs, Cys–Cys and Met–Met are among the most strongly represented. The preference for the sulfur-containing residues might reflect favorable packing interactions among the polarizable sulfur groups and, in the case of Cys may reflect the structural role of disulfide bridges in stabilizing folded β-sheet structures in peptides such as defensin HNP-3. Among heteropairs, Cys–Trp is the most strongly favored, again perhaps reflecting favorable van der Waals interactions. While this explanation for the high relative frequencies of Cys–Cys and Met–Met antiparallel (i, j) homopairs is appealing, it is probably not sufficient. Many of the antiparallel (i, j) homopairs come from homodimers, and in antiparallel homodimers $P(XX)/P(X)P(X)$ should be higher for less frequent amino acids. Thus, it should be noted that the most frequent amino acids in Figure 4 tend to have the smallest relative homopair frequencies in Figure 5, while the least frequent amino acids in Figure 4 tend to have the largest relative homopair frequencies in Figure 5.

Also among antiparallel (i, j) pairs, polar amino acids pair preferentially with polar amino acids and non-polar amino acids pair preferentially with nonpolar amino acids. Antiparallel (i, j) pairing among negatively charged acidic residues (Asp and Glu) and positively charged basic residues (Lys and Arg) favors oppositely charged pairs (Asp–Lys, Asp–Arg, Glu–Lys, and Glu–Arg) over like-charged heteropairs (Asp–Glu, Lys–Arg). Among this group, however, homopairs (Asp–Asp, Glu–Glu, Lys–Lys, and Arg–Arg) are strongly favored over heteropairs, presumably because of the large number of homodimers, even in the redundancy-reduced data set.

We envision that these statistics might be leveraged to identify new proteins involved in ICBS (see Discussion) and guide the *de novo* design of peptides and proteins that participate in interchain β-sheet formation (see also Smith and Regan, 1995, 1997). In particular, statistical information could provide a starting point for *de novo* computational design methods that are now becoming successful for short, single-chain proteins (Kuhlman *et al*., 2003).

4 DISCUSSION

A database of interchain β-sheet interactions is now available with an easy to use web-based interface. A ranking value (the ICBS index) has been developed to quantify the relative contributions of the β-ladders in the overall interchain interaction. Analysis of the database reveals a 15.8% prevalence of significant ICBS interactions, the majority of which involve the formation of antiparallel β-sheets and many of which involve the formation of dimers and oligomers. The frequencies of amino acids in ICBS interfaces are similar to those in intrachain β-sheet interfaces; antiparallel (i, j) homopairs are especially prevalent due to the high frequency of antiparallel homodimers. Examination of the ICBS interactions at a molecular level reveals a full range of non-covalent interactions that complement the hydrogen bonds that characterize the β-sheet interactions, including salt bridges, hydrogen bonds, hydrophobic interactions, and van der Waals interactions. Among antiparallel (i, j) pairs, polar amino acids pair preferentially with polar amino acids and non-polar amino acids pair preferentially with non-polar amino acids.

Statistical analysis of the amino acid composition of different types of interfaces, in particular within and between protein chains, are not always consistent in the literature with reports of both similarity (Keskin *et al*., 1998; Glaser *et al*., 2001; Jones *et al*., 2000) and differences (Jones and Thornton, 1997; Conte *et al*., 1999; Ofran and Rost, 2003). Our result, i.e. the first-order composition of ICBS strands is similar to that of intrachain strands is not necessarily in contradiction with the largest and most comprehensive study reported in (Ofran and Rost, 2003) for multiple reasons. In particular, Ofran and Rost do not compute statistics according to secondary structure, but

rather include other features, such as the transient/permanent nature of the interface.

The current version of the ICBS database should be viewed only as a first step towards the systematic study of ICBS interactions that can be extended in several directions. The ICBS index, for instance, may be complemented in time with more precise, but more complex, energetic calculations. In turn, these calculations may be able to shed light on other difficult questions that have been left out the present version of the database, such as the issue of conformational changes during the formation of ICBS complexes.

The statistics may also be expanded and refined by including additional properties such as edge versus central strands and/or degree of solvent exposure. As databases grow, better statistics may also be extracted allowing the comparison, for instance, of second-order statistics for edge and central strands in intrachain and interchain β-sheet interactions. In particular, these statistics may shed light on the tradeoffs between specific molecular recognition and avoidance of protein aggregation for exposed edge strands which are currently not entirely understood (Richardson and Richardson, 2002; Thirumalai *et al*., 2003). Another potentially interesting distinction that may be included in future versions of the ICBS database is the one between permanent and transient ICBS interfaces. While the corresponding information is currently absent from the PDB, it may be inferred with some degree of accuracy from the Swissprot database using the shortcuts developed in (Ofran and Rost, 2003).

We anticipate that the ICBS database and statistics derived from it, in combination with other databases and tools, might be leveraged to identify new ICBS interactions. For a given chain, the strongest criteria for ICBS interaction is proper homology with a sequence in the ICBS database. However, combinations of weaker criteria can also be used to draw ICBS inferences, including: (a) homology to sequences in one of the rapidly growing databases of protein–protein interactions (Xenarios and Eisenberg, 2001), such as DIP (Xenarios *et al*., 2002) and GRID http://biodata.mshri.on.ca/grid/; (b) presence of exposed β-strands in the chain; and (c) good agreement with ICBS first- and second-order statistics. When a protein has no close homologue in the PDB, secondary structure and relative solvent accessibility prediction programs, derived by machine learning methods, can be used to identify putative exposed β-strands. While not perfect, such programs achieve performances in the 75–80% range at the level of single amino acid (Pollastri *et al*., 2001a,b). In addition, these programs could be combined with programs for predicting edge strands (Siepen *et al*., 2003) to further improve detection. In a similar vein, current protein–protein interaction databases are notoriously noisy but are bound to improve over time. How to derive rational inferences from different combinations of predicted structural features in combination with noisy information from the protein–protein interaction databases and ICBS first- and second-order statistics is the object of ongoing research. But in the future, one ought to be able to infer whether a given chain/strand is involved in ICBS interactions and, if that is the case, what are the sequences and properties of its partners.

As protein–protein interactions are beginning to emerge as targets for drug development, the ICBS database may thus help identify or predict new targets for these efforts and *de novo* design peptides, proteins, or other compounds that participate or control ICBS formation. We are now using the data derived from analysis of the ICBS database to gain fundamental insights into the nature of ICBS interactions and to guide the development of chemical compounds and drugs that modulate, mediate, or block these interactions (Nowick *et al*., 2000, 2002).

ACKNOWLEDGEMENTS

The authors thank the UC Systemwide Biotechnology Research and Education Program (UC BREP) and the UCI Institute for Genomics and Bioinformatics for support. P.F.B. thanks UCI for a Laurel Wilkening Faculty Award, Sun Microsystems, and the National Institutes of Health for additional support (LM-07443-01). J.S.N. thanks the Camille and Henry Dreyfus Foundation for a Camille Dreyfus Teacher-Scholar Award, the American Chemical Society for an Arthur C. Cope Scholar Award, and the National Institutes of Health for additional support (GM-49076).

REFERENCES

Baldwin,E.T., Weber,I.T., St Charles,R., Xuan,J.-C., Appella,E., Yamada,M., Matsushima,K., Edwards,B.F.P., Clore,G.M., Gronenborn,A.M. and Wlodawer,A. (1991) Crystal structure of interleukin 8: symbiosis of NMR and crystallography. *Proc. Natl Acad. Sci. USA*, **88**, 502–506.

Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Burrows,S.D., Doyle,M.L., Murphy,K.P., Franklin,S.G., White,J.R., Brooks,I., McNulty,D.E., Scott,M.O., Knutson,J.R., Porter,D., Young,P.R. and Hensley,P. (1994) Determination of the monomer–dimer equilibrium of interleukin-8 reveals it is a monomer at physiological concentrations. *Biochemistry*, **33**, 12741–12745.

Choudhury,D., Thompson,A., Stojanoff,V., Langermann,S., Pinkner,J., Hultgren,S.J. and Knight,S.D. (1999) X-ray structure of the FimC–FimH chaperone–adhesin complex from uropathogenic *Escherichia coli*. *Science*, **285**, 1061–1066.

Clore,G.M., Appella,E., Yamada,M., Matsushima,K. and Gronenborn,A.M. (1990) Three-dimensional structure of interleukin 8 in solution. *Biochemistry*, **29**, 1689–1696.

Conte,L.L., Chothia,C. and Janin,J. (1999) The atomic structure of protein–protein recognition sites. *J. Mol. Biol.*, **285**, 2177–2198.

Doyle,D.A., Lee,A., Lewis,J., Kim,E., Sheng,M. and MacKinnon,R. (1996) Crystal structures of a complexed and peptide-free membrane protein-binding domain: molecular basis of peptide recognition by PDZ. *Cell*, **85**, 1067–1076.

Glaser,F., Steiberg,D.M., Vakser,I.A. and Ben-Tal,N. (2001) Residue frequencies and pairing preferences at protein–protein interfaces. *Prot.: Struct. Funct. Genet.* **43**, 89–102.

Hammarström,P., Schneider,F. and Kelly,J.W. (2001) *Trans*-suppression of misfolding in an amyloid disease. *Science*, **293**, 2459–2462.

Henrick,K. and Thornton,J.M. (1998) PQS: a protein quaternary structure file server. *Trends Biochem. Sci.*, **23**, 358–361.

Hill,C.P., Yee,J., Selsted,M.E. and Eisenberg,D. (1991) Crystal structure of defensin HNP-3, an amphiphilic dimer: mechanisms of membrane permeabilization. *Science*, **251**, 1481–1485.

Huizinga,E.G., Tsuji,S., Romijn,R.A., Schiphorst,M.E., de Groot,P.G., Sixma,J.J. and Gros,P. (2002) Structures of glycoprotein Ibα and its complex with von Willebrand factor A1 domain. *Science*, **297**, 1176–1179.

Hutchinson,E.G., Sessions,R.B., Thornton,J.M. and Woolfson,D.N. (1998) Determinants of strand register in antiparallel β-sheets of proteins. *Protein Sci.*, **7**, 2287–2300.

Jacobs,S.A. and Khorasanizadeh,S. (2002) Structure of HP1 chromodomain bound to a lysine 9-methylated histone H3 tail. *Science*, **295**, 2080–2083.

Jenuwein,T. and Allis,C.D. (2001) Translating the histone code. *Science*, **293**, 1074–1080.

Jones,S., Marin,A. and Thornton,J.M. (2000) Protein domain interfaces: characterization and comparison with oligomeric protein interfaces. *Protein Eng.*, **13**, 77–82.

Jones,S. and Thornton,J.M. (1997) Analysis of protein-protein interactions sites using surface patches. *J. Mol. Biol.*, **272**, 121–132.

Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.

Keskin,O., Bahar,I., Badretinov,A.Y., Ptitsyn,O.B. and Jernigan,R.L. (1998) Empirical solvent-mediated potentials hold for both intramolecular and inter-molecular inter-residue interactions. *Protein Sci.*, **7**, 2578–2586.

Korchuganov,D.S., Nolde,S.B., Reibarkh,M.Y., Orekhov,V.Y., Schulga,A.A., Ermolyuk,Y.S., Kirpichnikov,M.P. and Arseniev,A.S. (2001) NMR study of monomer-dimer equilibrium of barstar in solution. *J. Am. Chem. Soc.*, **123**, 2068–2069.

Kraulis,P.J. (1991) MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Cryst.*, **24**, 946–950.

Kuhlman,B., Dantas,G., Ireton,G.C., Varani,G., Stoddard,B.L. and Baker,D. (2003) Design of a novel globular protein fold with atomic-level accuracy. *Science*, **302**, 1364–1368.

Maitra,S. and Nowick,J. (2000) β-sheet interactions between proteins. In Greenberg, A., Breneman, C. and Liebman, J. (eds), *The Amide Linkage: Structural Significance in Chemistry, Biochemistry and Material Science*. Wiley, New York, pp. 495–518.

Merritt,E.A. and Bacon,D.J. (1997) Raster3D: photorealistic molecular graphics. *Meth. Enzymol.*, **277**, 505–524.

Nassar,N., Horn,G., Herrmann,C., Scherer,A., McCormick,F. and Wittinghofer,A. (1995) The 2.2 Å crystal structure of the Ras-binding domain of the serine/threonine kinase c-Raf1 in complex with Rap1A and a GTP analogue. *Nature*, **375**, 554–560.

Nowick,J.S., Chung,D.M., Maitra,K., Maitra,S., Stigers,K.D. and Sun,Y. (2000) An unnatural amino acid that mimics a tripeptide β-strand and forms β-sheetlike hydrogen-bonded dimers. *J. Am. Chem. Soc.*, **122**, 7654–7661.

Nowick,J.S., Lam,K.S., Khasanova,T.V., Kemnitzer,W.E., Maitra,S., Mee,H.T. and Liu,R. (2002) An unnatural amino acid that induces β-sheet folding and interaction in peptides. *J. Am. Chem. Soc.*, **124**, 4972–4973.

Ofran,Y. and Rost,B. (2003) Analysing six types of protein–protein interfaces. *J. Mol. Biol.*, **325**, 377–387.

Pedersen,A.G., Baldi,P., Chauvin,Y. and Brunak,S. (1998) DNA structure in human RNA polymerase II promoters. *J. Mol. Biol.*, **281**, 663–673.

Petosa,C., Collier,R.J., Klimpel,K.R., Leppla,S.H. and Liddington,R.C. (1997) Crystal structure of the anthrax toxin protective antigen. *Nature*, **385**, 833–838.

Pollastri,G., Baldi,P., Fariselli,P. and Casadio,R. (2001a) Prediction of coordination number and relative solvent accessibility in proteins. *Proteins*, **47**, 142–153.

Pollastri,G., Przybylski,D., Rost,B. and Baldi,P. (2001b) Improving the prediction of protein secondary strucure in three and eight classes using recurrent neural networks and profiles. *Proteins*, **47**, 228–235.

Rafferty,J.B., Somers,W.S., Saint-Girons,I. and Phillips,S.E. (1989) Three-dimensional crystal structures of *Escherichia coli* met repressor with and without corepressor. *Nature*, **341**, 705–710.

Rajarathnam,K., Sykes,B.D., Kay,C.M., Dewald,B., Geiser,T., Baggiolini,M. and Clark-Lewis,I. (1994) Neutrophil activation by monomeric interleukin-8. *Science*, **264**, 90–92.

Ratnaparkhi,G.S., Ramachandran,S., Udgaonkar,J.B. and Varadarajan,R. (1998) Discrepancies between the NMR and X-ray structures of uncomplexed barstar: analysis suggests that packing densities of protein structures determined by NMR are unreliable. *Biochemistry*, **37**, 6958–6966.

Richardson,J.S. and Richardson,D.C. (2002) Natural β-sheet proteins use negative design to avoid edge-to-edge aggregation. *Proc. Natl Acad. Sci. USA*, **99**, 2754–2759.

Sauer,F.G., Füterer,K., Pinkner,J.S., Dodson,K.W., Hultgren,S.J. and Waksman,G. (1999) Structural basis of chaperone function and pilus biogenesis. *Science*, **285**, 1058–1061.

Siepen,J.A., Radford,S.E. and Westhead,D.R. (2003) β Edge strands in protein structure prediction and aggregation. *Protein Sci.*, **12**, 2348–2359.

Smith,C.K. and Regan,L. (1995) Guidelines for protein design: The energetics of β-sheet side chain interactions. *Science*, **270**, 980–982.

Smith,C.K. and Regan,L. (1997) Construction and design of β-sheets. *Acc. Chem. Res.*, **30**, 153–161.

Thirumalai,D., Klimov,D.K. and Dima,R.I. (2003) Emerging ideas on the molecular basis of protein and peptide aggregation. *Curr. Opin. Struct. Biol.*, **13**, 146–159.

Westbrook,J., Feng,Z., Jain,S., Bhat,T.N., Thanki,N., Ravichandran,V., Gilliland,G.L., Bluhm,W., Weissig,H., Greer,D.S., Bourne,P.E. and Berman,H.M. (2002) The Protein Data Bank: unifying the archive. *Nucleic Acids Res*, **30**, 245–248.

Wlodawer,A., Miller,M., Jaskolski,M., Sathyanarayana,B.K., Baldwin,E., Weber,I.T., Selk,L.M., Clawson,L., Schneider,J. and Kent,S.B. (1989) Conserved folding in retroviral proteases: crystal structure of a synthetic HIV-1 protease. *Science*, **245**, 616–621.

Xenarios,I. and Eisenberg,D. (2001) Protein interaction databases. *Curr. Opin. Biotechnol.*, **12**, 334–339.

Xenarios,I., Salwinski,L., Duan,X.J., Higney,P., Kim,S. and Eisenberg,D. (2002) DIP: The Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.*, **30**, 303–305.

APPENDIX: ADDITIONAL TECHNICAL CONSIDERATIONS

To limit the impacts of the format limitations, variations or errors linked to PDB and PQS files, we adopted a number of methodological choices. These choices will most likely become unnecessary when PDB and PQS entries are released in a different format or made accessible through direct database queries. We report them here for completeness.

Redundancy of PQS entries Likely protein quaternary structures, i.e. PQS entries, sometimes contain the same ICBS interfaces as the PDB structure they are derived from, and are in this respect redundant. To let users filter out such structures when querying the ICBS database, we use a simple but effective, empirical redundancy criterion. An ICBS entry corresponding to a PQS structure is considered redundant if one of the following conditions is met: (a) it represents only a part of the PDB structure; or (b) the proportion of chain pairs that display interchain β-ladders is the same in the PQS and in the PDB structures.

Identification of protein chains PDB files and the DSSP output format both use a single character to label protein or ligand chains. In the rare case of very large structures where the number of chains is such that the representation capabilities of PDB and DSSP labels are exhausted, and where some chains are therefore not uniquely identified through their label, we consider only the first chain with any given label. As in such cases the structure consists of numerous copies of the same PDB asymmetric unit, with repetitive ICBS interfaces between them, this solution is unlikely to cause any unique ICBS interaction to be missed.

Identification of β-bridges, ladders and β-sheets The DSSP program labels β-bridges with a single letter. Case is used to denote the parallel or antiparallel nature of the bridge. Only 26 unique identifiers are thus available in each category. β-sheets are labeled with capital letters. Theoretically, one would therefore have to rebuild the whole connectivity patterns of β-bridges and ladders to uniquely identify bridges, ladders and β-sheets within large proteins. Instead of adopting this costly solution, we forge an ICBS ladder label that is very unlikely to be non-unique, by combining the DSSP ladder label, the label of the β-sheet it belongs to, and the labels of the two pairing chains. We are thus able to count ladders and hydrogen-bonds, and to determine the interface orientation for each pair of chains.

Treatment of heteroatom PDB records Heteroatom PDB records (i.e. HETATM rows in PDB files) specify 'non-standard' groups. They sometimes contain amino acid residues that are an integral part of the protein and participate in interchain β-bridges. This is for instance the case for the quaternary structure of an HIV-1 protease (3hvp). Heteroatom records should therefore be taken into account in spotting and characterizing ICBS interactions. However, numerous variations in the way HETATM rows are used and placed in PDB files cause a number of undesirable effects. Until mmCIF or direct database queries are officially available in place of PDB files, we therefore ignore β-bridges and atom contacts involving a residue specified in a heteroatom row. As a consequence, we might under- or over-estimate the ICBS index for a few ICBS entries, and we might miss a few proteins in which ICBS interactions would only occur between residues specified in HETATM rows.

Case of very large entries that break the DSSP format In some rare cases, the DSSP sequential number that identifies residues can exceed the limit (9999) that the DSSP output format allows. β-Bridges involving such residues must in this case be ignored. As a consequence, we might occasionally miss some interchain β-bridges in very large proteins.

Case of inconsistent DSSP secondary structure assignment for β*-bridge partners* In a few cases, the output of the DSSP program specifies that a residue R1 pairs with another residue R2 through a β-bridge, whereas R2 is not found to pair with R1. We chose to consider that a β-bridge existed in such cases.