BIOINFORMATICS ORIGINAL PAPER Vol. 25 no. 18 2009, pages 2418–2424
doi:10.1093/bioinformatics/btp424

Databases and ontologies

ADAN: a database for prediction of protein–protein interaction of modular domains mediated by linear motifs

J. A. Encinar[1,†], G. Fernandez-Ballester[1,†,*], I. E. Sánchez[2], E. Hurtado-Gomez[1,3], F. Stricher[3], P. Beltrao[4] and L. Serrano[3,‡]

[1]Instituto de Biologıa Molecular y Celular, Edificio Torregaitan, Universidad Miguel Hernandez, 03202 Elche (Alicante), Spain, [2]Fundación Instituto Leloir and IIBBA-CONICET, Patricias Argentinas 435, C1405BWE Buenos Aires, Argentina, [3]EMBL/CRG Systems Biology Research Unit, Centre for Genomic Regulation (CRG), UPF, 08003 Barcelona, Spain and [4]Department of Cellular and Molecular Pharmacology, University of California, San Francisco, CA 94158, USA

Received on February 13, 2009; revised on July 6, 2009; accepted on July 7, 2009

Advance Access publication July 14, 2009

Associate Editor: Burkhard Rost

ABSTRACT

Motivation: Most of the structures and functions of proteome globular domains are yet unknown. We can use high-resolution structures from different modular domains in combination with automatic protein design algorithms to predict genome-wide potential interactions of a protein. ADAN database and related web tools are online resources for the predictive analysis of ligand–domain complexes. ADAN database is a collection of different modular protein domains (SH2, SH3, PDZ, WW, etc.). It contains 3505 entries with extensive structural and functional information available, manually integrated, curated and annotated with cross-references to other databases, biochemical and thermodynamical data, simplified coordinate files, sequence files and alignments. Prediadan, a subset of ADAN database, offers position-specific scoring matrices for protein–protein interactions, calculated by FoldX, and predictions of optimum ligands and putative binding partners. Users can also scan a query sequence against selected matrices, or improve a ligand–domain interaction.

Availability: ADAN is accessible at http://adan-embl.ibmc.umh.es/ or http://adan.crg.es/.

Contact: gregorio@umh.es

1 INTRODUCTION

The annotation of interaction domains into a database and the prediction of putative protein–protein interactions are important steps for the computational characterization of protein function at genomic scale. Proteins interact with other proteins to achieve their functions. The importance of these interactions in the cell organization is reflected in the fact that almost 80% of proteins interact with other partners (Gavin *et al*., 2006). Computational annotation of protein function is traditionally obtained through sequence similarity, where the identification of a protein function is automatically ascribed to homologous sequences. However, this approximation fails when the sequences diverge or when no close neighbors with known functions are available. In this case, the characterization of protein function can be approached on a structural basis (Fernandez-Ballester and Serrano, 2006), using the structures that are homo- and heteromeric protein complexes to understand the basis of protein interactions (Aloy and Russell, 2002) and validate interactions determined by other methods. Current databases of structure-based predictions of protein–protein interactions do not fully exploit the potential of recent developments in the field, such as the prediction of which sequences can be accommodated in a given interface (Fernandez-Ballester and Serrano, 2006) or the combined use of structural and biological information for the prediction of in vivo interactions (Sanchez *et al*., 2008). For example, the PRISM database includes exhaustive predictions of interactions between globular domains, but only for domains of known structure (Ogmen *et al*., 2005). Moreover, it does not link to available biochemical information for the known complexes. On the other hand, MODBASE is a comprehensive database for homology models of globular domains but does not include predicted interactions (Pieper *et al*., 2006). Although state-of-the-art methods are able to predict interactions between globular domains and short linear motifs (Sanchez *et al*., 2008), neither PRISM nor MODBASE make predictions for this important class of complexes.

Genetically mobile domains are structural/functional units that appear in protein architectures. The set of these domains comprises a few hundreds of families that are easily recognized, classified and organized in useful databases such as SMART (Letunic *et al*., 2006; Ponting *et al*., 1999; Schultz *et al*., 1998). Taking advantage of this growing structural information, we have collected domain structures, in complex with polypeptide ligands, when available, in a database for the prediction of protein–protein interActions of moDular domAiNs, mediated by linear motifs (ADAN). This database organizes and links a large number of protein domain structures annotated with biochemical and functional data and binding predictions, thus providing a molecular picture of proteins and their interactions. ADAN database represents a launching platform for studies on protein–protein interactions in general because it facilitates template selection for homology modeling

*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.

‡L.S. is a ICREA professor

of neighbor sequences, and selection of high-resolution ligands to construct ligand–domain complexes. It also facilitates the analysis of models and structures in terms of theoretical energy by means of FoldX (Guerois *et al*., 2002; Schymkowitz *et al*., 2005) through the construction of scoring matrices and prediction of optimum ligands and putative partners at genome level. Currently, ADAN contains 3505 entries, corresponding to high-resolution X-ray (42.6%), NMR (15.8%) and modeled structures (41.6%).

2 METHODS

2.1 Domain isolation and classification

Domains included at present in ADAN comprise 23 different families, representing 15% of the known signaling domains included in SMART database. These domains are mainly involved in interaction with peptides or proteins: 14–3–3, BRCT, FHA, PDZ, PH, Polo box, PTB, PTPc, RA, RBD, SH2, SH3, UBQ, VHS, WD40, WW, ARM, FF, MH2, TRP, but we also include catalytic domains, such as methyl transferases, phosphoserine phosphatases and kinase superfamily. Family selection for inclusion in ADAN depends on structure abundance (i.e. there are 430 SH3, 191 SH2, 289 PDZ domains, etc.), relevance and interest of the domain (i.e. kinases, methyl transferases). Other families can be included when more data become available.

The entries in ADAN database are simplified structural coordinate files from the PDB database. The structural files from Protein Data Bank are never used directly because protein complexes often contain multiple polypeptide chains, which in turn contain several identical or different domains. The coordinate files having an interacting domain are individually extracted, including its interacting peptide ligands, cofactors and metal ions in active sites. Water and other extra molecules and long C- and N-terminal domain extensions are cut out when possible without interfering either with domain-folding architecture or ligand-binding area. Coordinate files with one or several peptide chains containing two or more globular domains are extracted separately and treated as individual structures. When available, intrachain ligands (i.e. SH2 and SH3 in whole Src-kinases) are isolated in complex with the interacting domain. Finally, coordinate files having identical repeats in the asymmetric unit are checked for structure integrity (broken loops, etc), and the best repeat selected. The simplified coordinate files are systematically named with the original PDB code name followed by a number (starting by 2): as an example the PDB entry 1OV3.PDB corresponds to P47phox human protein and contains two SH3 domains. The entry names in ADAN were 1OV3**2**.PDB and 1OV3**3**.PDB for the two SH3 domains.

Two groups of structures are usually differentiated: simple modular protein domains, and modular protein domains in complex with peptidic ligands. All these domain structures are directly introduced in ADAN as single entries, annotated with available information and stored ready for future use as templates (i.e. for homology modeling). Structures containing valid domain–ligand interactions are selected and used for predictions (see below).

Models by homology have been constructed in some cases to expand the set of structures in complex with ligands as described previously (Marti-Renom *et al*., 2000). The modeling requires the identification of the right template using information derived from the comparison of sequences and structures (Fernandez-Ballester and Serrano, 2006; Fernandez-Ballester *et al*., 2009), and a quality assessment evaluation in terms of energy (see below).

2.2 Energy evaluation, mutagenesis and interaction scoring matrices

Structural energy analysis and mutagenesis were conducted using the FoldX algorithm (Guerois *et al*., 2002; Schymkowitz *et al*., 2005). The use of FoldX in protein design has been extensively reported (Fernandez-Ballester *et al*., 2004; Kempkens *et al*., 2006; Kiel and Serrano, 2006; Kiel *et al*., 2004, 2005; Kolsch *et al*., 2007; Sanchez *et al*., 2008; Villanueva *et al*., 2003). Briefly, FoldX is a force field developed for the rapid evaluation, stability, folding, and dynamics of proteins to assess the effect of mutations. The algorithm provides a fast and quantitative estimation of the interactions contributing to the stability of proteins and protein complexes. The different energy terms taken into account in FoldX have been weighed using empirical data from protein engineering experiments, and the predictive power has been tested on a very large set of protein mutants, covering most of the structural environments found in proteins. In the development of ADAN, FoldX was used mainly to repair structures prior to analysis and to perform mutagenesis and evaluation of interaction energy (Fernandez-Ballester and Serrano, 2006; Fernandez-Ballester *et al*., 2004). FoldX also aided in the generation of models by homology based on sequence alignments, superimposition of structures and creation of chimeras, and for docking of selected ligands on structure/models. In depth details are described in previous publications (Guerois *et al*., 2002; Kiel *et al*., 2004; Schymkowitz *et al*., 2005).

Mutagenesis was performed by FoldX using the BuildModel and PositionScan commands. These procedures test different rotamers and allow neighbor side-chains to move. In this way, we ensure that mutations and energy values obtained for a given position are not obscured by neighbor positions. BuildModel was used to mutate the all amino acids in the peptidic ligands to poly-alanine, with the exception of Gly and phosphorylated residues, to obtain the starting template for the positional mutagenesis. PositionScan was used to mutate each position to 20 natural amino acids. For domains known to bind phospho peptides (such as the SH2 domain), the natural amino acids were complemented with the phosphorylated ones (pSer, pThr and pTyr) during the mutagenesis of the target peptide (Sanchez *et al*., 2008). Matrices are 2D tables containing the interaction energy obtained for all residues built in all ligand positions. Each position in the ligand is mutated individually, while the other positions remain as poly-Ala. The interaction energy is considered as a score that represents the relative importance of each amino acid at each position. The matrices are corrected by adding internal van der Waals clashes of the interface residues with their own chains to the binding energy, and normalized to the lower value (becoming 0). The lower the score value, the better the ligand–domain interaction. For a given binding matrix, the binding score of a sequence can be calculated by summing over all positions of the matrix, which are taken to be independent.

2.3 Additional genomic information used in ADAN

2.3.1 Genome sequences FASTA formatted files containing the entire genome of species were downloaded from Protein Knowledgebase UniprotKB: (http://beta.uniprot.org/downloads/).

Species included in the genome scanning were: *Acanthamoeba castellanii*, *Ashbya gossypii*, *Arabidopsis thaliana*, *Avian sarcoma virus*, *Bacillus subtilis*, *Bos taurus*, *Candida albicans*, *Caenorhabditis elegans*, *Canis familiaris*, *Cryptococcus neoformans*, *Cryptosporidium parvum*, *Danio rerio*, *Dictyostelium discoideum*, *Drosophila melanogaster*, *Escherichia coli*, *Gallus gallus*, *Human immunodeficiency virus*, *Homo sapiens*, *Methanococcus jannaschii*, *Methanobacterium thermoformicicum*, *Murine cytomegalovirus*, *Mus musculus*, *Mycobacterium tuberculosis*, *Nicotiana tabacum*, *Oryctolagus cuniculus*, *Plasmodium falciparum*, *Pseudomonas aeruginosa*, *Pyrococcus horikoshii*, *Rattus norvegicus*, *Rous sarcoma virus*, *Saccharomyces cerevisiae*, *Salmonella typhimurium*, *Scenedesmus obliquus*, *Schizosaccharomyces pombe*, *Streptococcus pneumoniae*, *Sus scrofa*, *Thermus filiformis*, *Thermus thermophilus*, *Tobacco etch virus*, *Xenopus laevis*, *Yersinia enterocolitica*, *Yersinia pestis* and *Zea mays*.

2.3.2 Protein information All entries in ADAN database are used to interrogate Expasy database (http://www.expasy.ch/) for the following fields: SwissProt code, gene names and synonyms, gene locus and open reading frames, organism, taxonomy, function, location, interactions, post-translational modifications, cross-references, and PubMed references. The protein location information obtained is grouped into several compartments

due to the great dispersion of terms used for describing protein location. As an example, the terms 'er membrane', 'endoplasmic reticulum' and 'endoplasmic reticulum membrane' were grouped, and proteins having one of these terms were treated similarly. Plasma membrane proteins were considered as cytoplasmic proteins since their N- and/or C-terminal are usually cytosol-exposed and interact with soluble proteins. Non-classified localization was treated as cytoplasmic as well. No distinction was made for sub-subcellular locations, as mitochondrion inner or outer membrane. The entire genome of each species was pre-calculated for globularity with GlobPlot (Linding *et al*., 2003) (http://globplot.embl.de) for use in genome scanning. The patterns used for genome scanning were the less restrictive possible: PxxP or PxxxP for SH3, Yx (at least one Tyr in the peptide) for SH2, [ST] × [IVLA] or [IVLAMFYW] × [IVLA] (C-terminal peptide) for PDZ, and Rxx[ST] × [PG] for 14-3-3. The information on protein–protein interactions for all species was downloaded from MINT ftp service: (ftp://mint.bio.uniroma2.it/pub/release/txt/current/).

2.4 Database maintenance

Database maintenance and update were automated through the development of several Python (http://www.python.org) scripts to detect new incoming structures, the corresponding links to other databases and the remaining information to complete annotation. Nevertheless, the new database entries were manually curated by a scientific team to confirm the identity of the individual domains in multi-complex structures and to isolate and prepare final ADAN entries for further predictive processing. The ADAN web is hosted on a Windows Server 2003 Enterprise Edition. A search engine has been developed in ASP (Server Activated Pages) code to query the database and WEB display. The webpage allows the visualization and data search in host server at http://adan-embl.ibmc.umh.es/ (IBMC-UMH, Elche, Spain) and the mirror http://adan.crg.es/ (CRG, Barcelona, Spain).

3 RESULTS

ADAN comprises two main modules: the database that holds modular domains and their annotation, and PREDIADAN, a subset of the ADAN database containing pre-calculated predictive data, and tools for predictive analysis of protein complexes.

3.1 Database and web server description

ADAN database for modular domains is composed of entries selected from structures, or derived models belonging to 23 domain families (see 'Methods' section), and annotated with functional and biochemical data. The main page of the ADAN web server shows some records with part of the annotation fields, including a quick search tool that allows for general queries (Fig. 1a). This quick search is, however, limited to those fields marked with an asterisk. An advanced search tool allows the user to search a broad range of fields and Boolean combinations: domain name, protein family, references, source organism, etc (Fig. 1c). It should be noted that the current version of ADAN doesn't include information of gene names or genes ids, and so the search will not produce results. A summary is available for each domain entry in an abstract (Fig. 1b), containing whole protein information and related databases [SwissProt (Gasteiger *et al*., 2001), Interpro (Apweiler *et al*., 2001; Mulder *et al*., 2005), Pfam (Finn *et al*., 2006), Uniprot (Bairoch *et al*., 2005)], pre-calculated alignment for whole protein or isolated domains, linked to the multiple sequence alignment server SAS (Milburn *et al*., 1998), and modular domain information with protein description annotation, structure resolution, source organism, thermodynamic data, PubMed references, etc., linked to Pfam (Finn *et al*., 2006), RCSB_PDB (Berman *et al*., 2003), SMART

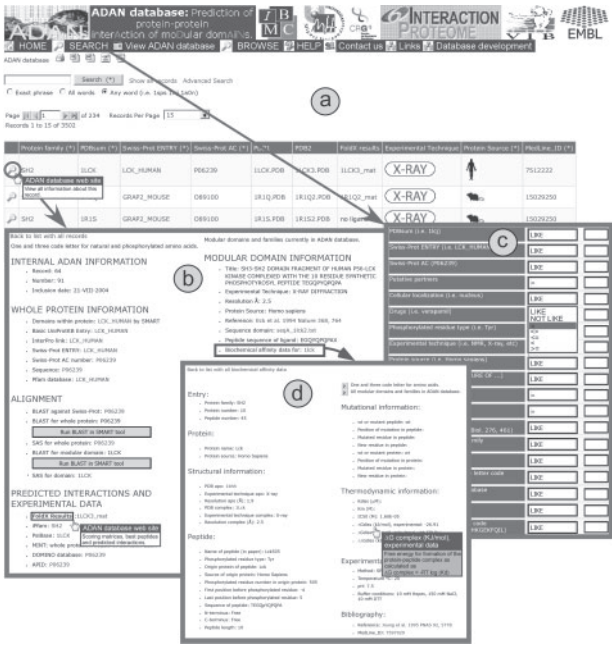

Fig. 1. Main module of ADAN database. The main page of ADAN shows some of the annotations included in the database, such as protein family, links to other databases, links to coordinate files, protein source, etc. (**a**) The illustration shows the results of a quick search using SH3 as query text. The dynamic page generated offers the possibility to browse the results page-by-page, 15 records each by default (or other selected by the user). (**b**) Magnifying glasses in each record connect to the abstract, where all annotated information is presented in blocks. Links connect not only to external, but also to ADAN generated information. (**c**) An advanced search engine, comprising most of the annotated fields, allows complex searches in the database. (**d**) Biochemical and thermodynamic affinity data are stored in a subset of the database containing the list of peptide-domain interactions compiled from bibliography. Each record in ADAN has a direct link to these data (if available), where annotation is presented in blocks. The page can also be searched in the same way than the main module of ADAN and is directly accessible from internet (http://adan-embl.ibmc.umh.es/thermo.asp). All pages contain self-explanatory pop-up windows to guide users. Screenshots in all figures are taken from the SH2 domain of the human tyrosine kinase 1LCK.PDB.

(Ponting *et al*., 1999; Schultz *et al*., 1998), PDBsum (Laskowski *et al*., 2005), and MSD (Tagari *et al*., 2006).

A fraction of ADAN entries are annotated with biochemical and thermodynamic data. When available, the database shows values for K_d, K_m, IC_{50}, ΔG of the complex, $\Delta\Delta G$ of dissociation, buffer conditions and PubMed reference for one or several peptides interacting with the ADAN entry, grouping all this information in blocks (Fig. 1d). Currently there are a total of 1925 annotations for a small subset of ADAN entries belonging to 16 domain families (14_3_3; BRCT; FHA; Histone acetyltransferase; PDZ; Phosphoserine phosphatase; Polo box; Tyrosine phosphatase; PTB; RA; RBD; SH2; UBQ; VHS; WD40 and WW) that interact with (or are related to) phospho peptides. The most represented set of interacting phospho peptides is for SH2, which contains 1224 interacting peptides belonging to 36 different SH2 entries in ADAN. The ADAN dataset containing biochemical and thermodynamical data can be accessed directly through http://adan-embl.ibmc.umh.es/thermo.asp.

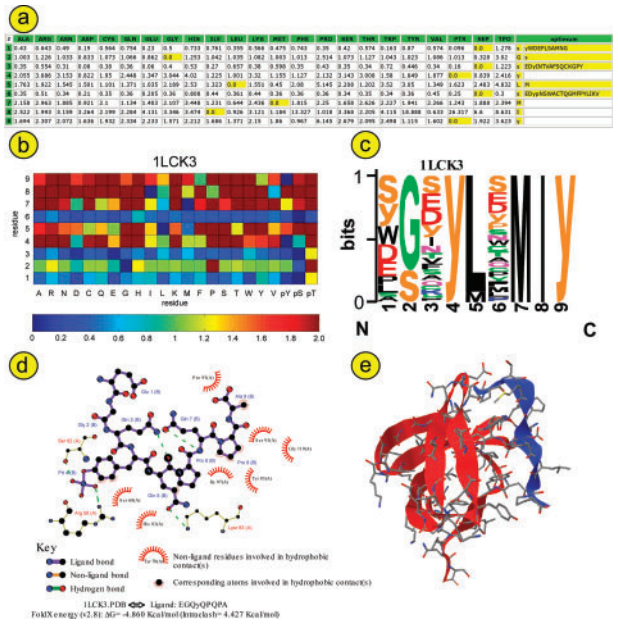

Fig. 2. All entries in ADAN in complex with a peptidic ligand have pre-calculated predictions of protein interaction. (**a**) Position specific scoring matrices. The matrices are calculated position by position by means of FoldX. The algorithm mutates each position in the ligand individually to 20 natural amino acids, and evaluates the energy to construct the table. Values are normalized, lower values meaning better binding interactions (0 is the best). (**b**) Color code representation of the scoring matrix. Values in the matrix in the range 0–2 kcal/mol are transformed into a color scale and plotted. Blue color indicates good binding interactions, and the red color, the contrary. (**c**) Graphical interpretation of the scoring matrix with the Alpro program. Letters in the graph were plotted when the energy value is within the 0.5 kcal/mol range. Residues with better binding energies are plotted in a larger size. (**d**) Schematic representation of the wild-type ligand derived from the template used for matrix calculation. Graph was prepared with LigPlot and includes the wild-type binding energy to be used as reference. (**e**) 3D visualization of the molecules. Simplified structures and models can be viewed through Jmol applet (http://www.jmol.org/) to assess the reliability of the scoring matrices.

3.2 Prediadan

The ADAN module for prediction of protein-protein interactions includes at present 386 X-ray and 102 NMR domain–ligand complexes directly identified and selected from the PDB without any modeling procedure. In addition, there are 1423 models made for yeast SH3 (Fernandez-Ballester *et al*., 2009), covering 21 out of 27 *S.cerevisiae* SH3, and 29 Ras-Rab models (Kiel *et al*., 2004). The specificity and interaction prediction information available for each complex is presented in a separate page, available through links in the main module of the database under predicted interactions and experimental data (Fig. 1b). It derives from the calculation performed with FoldX on the complex structure (see 'Methods' section), and comprises a scoring matrix that describes the ability of the 20 natural amino acids to fit in a position of the ligand in a complex (Fig. 2), being the best peptides to act as ligands and putative protein partners detected by genome scanning. The page also contains links to related databases dealing with protein interactions: iPfam (Finn *et al*., 2005), PsiBase (Gong *et al*., 2005a, b), MINT (Chatr-aryamontri *et al*., 2007), DOMINO (Ceol *et al*., 2007) and APID (Prieto and de las Rivas, 2006).

3.2.1 Position-specific scoring matrices Scoring matrices were calculated from crystallographic or modeled structures, in complex with poly-Ala ligands by means of FoldX (see 'Methods' section). Each position in the ligand was explored individually in the sequence space and the neighboring positions in the domain were relaxed to avoid clashes (Fernandez-Ballester and Serrano, 2006; Schymkowitz *et al*., 2005). The resulting structures were evaluated in terms of energy, allowing the selection of the better residue(s) per position. The result is a scoring matrix of normalized binding energies (kcal/mol) that reflect the ability of an amino acid in a ligand position to fit in the domain under study (Fig. 2a). For each entry in Prediadan we show the normalized matrix, and the un-normalized binding and stability matrices. The optimum amino acid in each position is highlighted in yellow. In addition, the residues within the range of 0.5 kcal/mol with respect to the best are sorted, written, and highlighted.

The matrix is transformed into color code graphics with MatLab (http://www.mathworks.com) (Fig. 2b), ranging from 0 (best value, dark blue) to 2 kcal/mol (worst value, dark red). Values higher than 2 kcal/mol are displayed in dark red. The matrix is transformed into a graphic with Alpro (http://www.ccrnp. ncifcrf.gov/∼toms/) (Fig. 2c). The letters are rescaled to 1 and sized according to their values in the matrix. To simplify, we show only those residues within 0.5 kcal/mol of the energy of the best amino acid. As a reference, the wild-type ligand sequence appearing in the original structure and their binding energy (calculated by FoldX) are schematically drawn (Fig. 2d) by LigPlot (Wallace *et al*., 1995). In addition, ligand–domain structures can be viewed in 3D through Jmol (Fig. 2e) to assess the reliability of the scoring matrices.

3.2.2 Optimum ligands Each entry in Prediadan is linked to a list of the best ligands derived from the scoring matrix. We extracted from each matrix the best residues per position (within a threshold 0.5 kcal/mol, and up to a maximum of three amino acids) and constructed all possible combinations to get a set of putative ligands. These ligands are scored using the matrix. The top 90 best ligands are shown, and the top 30 putative ligands are modeled and linked in ADAN database. The construction of the best 30 putative ligands was accomplished with FoldX, based on the poly-Ala ligand templates, and mutating all the positions at the same time. The binding energy was computed on the generated complexes and again sorted, looking at intraclashes (using the strongest van der Waals parameters). The pre-computed data (Fig. 3c) is presented in the web page along with the energy values of the wild-type templates used for mutagenesis as reference (Fig. 3a and b), and suggested as putative binders. The user has also the opportunity to select other non-precalculated putative ligands and send them to the web server for calculations.

3.2.3 Genome scan The ADAN genome scan tool predicts putative targets of a peptide binding domain through proteome wide analysis using available information (i.e. protein sequence, localization, interactions, etc). Proteins that do not share the same cellular compartment as the query domain are discarded. All possible putative ligand peptides from the remaining protein sequences are scored with the corresponding matrix. These peptide lists are filtered using several criteria: (i) *Random-threshold*: We defined a random-threshold for each matrix as the average binding energy of a pool of peptide fragments derived from all the proteins included in the corresponding genome. Peptides with score above this threshold

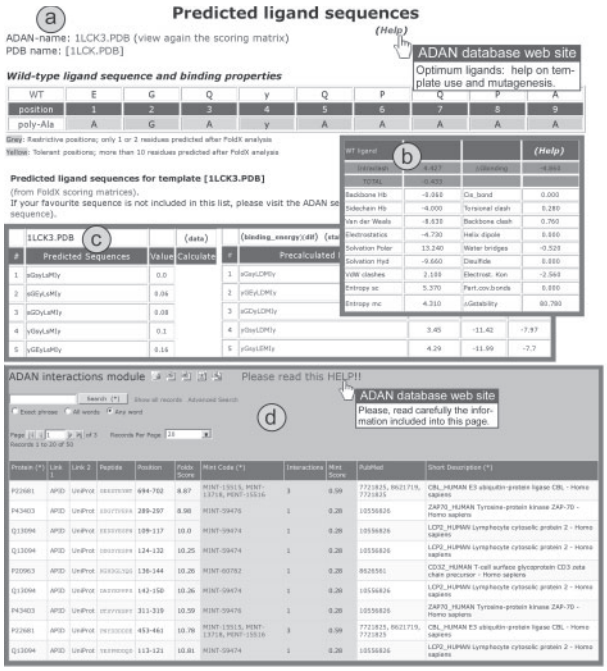

Fig. 3. Calculation of potential peptidic ligands from scoring matrices. (**a**) The web page shows the ligand–domain complex under study (i.e. the SH2 domain 1LCK3.PDB), and the results obtained after analysis of its scoring matrix. The wild-type ligand and the corresponding poly-Ala ligand sequence used for calculations are detailed. (**b**) Decomposition of the binding energy of the wild-type template under study. These values serve as reference to compare results after remodeling the template with the best results. (**c**) Using the scoring matrices, the three best amino acids per position are selected, combined and ranked in a list with the best putative binders. Only the best 90 peptides generated and their values are presented (left column). The best 30 peptides are remodeled on the template with FoldX, evaluated by energy and sorted by the arithmetic sum of binding energy plus intra-clashes (right column). The remaining 60 best peptides were not pre-computed, but a selection can be sent to the server on demand. (**d**) Interactor module: genome scans for an ADAN record in complex with a peptidic ligand. The information is stored in a searchable web page containing the list of hits obtained after genome scan, and filtering of putative partners, as described in 'Methods' section. The fields shown include the interacting protein name, links of these proteins with other databases, peptide sequence and position, FoldX score, MINT information, etc. The MINT code field is shown as 'nd' when the hit is found as favorable with FoldX, but no information is available in MINT. The page includes quick and advanced search tools, as well as a useful help for users.

are discarded; (ii) *Globularity*: Most ligands in the domain–ligand complex structures studied here have disordered conformations. For this reason, we select only peptides that are predicted to be 'disordered' and 'not globular' at the same time; (iii) *Pattern*: General patterns described in the bibliography are used to select peptides; (iv) *MINT*: Surviving peptides are tested for already described interactions in the MINT protein interaction database (Chatr-aryamontri *et al*., 2007). In contrast to the other filtering steps, we did not discard any putative peptide target, since the interactions not experimentally described to date may still take place in nature.

Currently, the ADAN database contains predicted interaction information for 65 SH2, 80 SH3, 53 PDZ, 15 14-3-3 structures, and for the 1423 yeast SH3 models. For an easy access of this information we created a searchable web page containing the list of potential partners, and including whether or not the interactions were already described experimentally. Additional information includes peptide sequence and position within the protein, FoldX score, Mint code and score, PubMed code describing the interactions, and short descriptions of the interacting proteins with links to other databases. Figure 3d shows the ADAN web pages and link connections for an SH3 record in complex with a peptidic ligand.

3.2.4 Prediction from a query sequence This web-based tool in PREDIADAN offers the users the opportunity to analyze a sequence of their interest. A protein sequence of any length is scanned against the selected matrices to look for regions that could act as putative ligands. The sequence is analyzed with a sliding window of length equal to the length of the ligand in the structure/model selected. Each peptide read is evaluated with the position specific scoring matrices and the best 10 peptides are automatically modeled on the selected template(**s**). The user gets a compressed file by e-mail, containing the top scored regions within the sequence, the top 10 modeled structures, and the corresponding energy files. This tool is directly available online at http://adan-embl.ibmc.umh.es/prediadan.asp**.**

3.3 Benchmarking of the ADAN binding matrices

The predicted binding matrices stored in ADAN can be used as part of predictors for protein–protein interactions. It is important at this point to make the distinction between *in vitro* domain-peptide interactions and the *in vivo* protein–protein interactions mediated by these same domains. The probability that two proteins interact in the cell is a function of more than the binding affinity of one of its domain against a peptide within a putative interaction partner (ex. gene expression, co-localization, secondary structure of the peptide, etc). Given that these matrices provided in ADAN predict the binding against target peptides these should be validated against *in vitro* peptide binding experiments. We have previously validated the predictive power of these matrices for two domain types for which a large amount of experimental peptide binding data is available. We analyzed six *S.cerevisiae* SH3 domains (Fernandez-Ballester *et al*., 2009) and seven human SH2 domains (Sanchez *et al*., 2008) using the area under the receiver operating characteristics (ROC) curve as a performance measure to test the capacity of the FoldX matrices to discriminate between experimentally determined binding and non-binding peptides. For the human SH2 domains we observed an average area under the ROC curve (AROC) of 0.68 while for the *S.cerevisiae* SH3 domains the average AROC value was 0.81. It is important to note that for the yeast SH3 domains we used different models of the same domain to improve the specificity predictions. We also caution that specific matrices might vary in their performance. We observed AROC values ranging from 0.44 (for the human Stat1 SH2 domain) to 0.96 (for the yeast SH3 domains of Lsb3).

For the SH3 domains of *S.cerevisiae* we additionally tested directly the correlation between predicted binding to peptides and experimentally determined dissociation constants (Fernandez-Ballester *et al*., 2009). For 45 peptides with available data we observed a significant correlation of 0.35 ($P < 0.02$), which shows us that these matrices can be used as semi-quantitative predictor of domain-peptide binding strength.

Although the main objective of the ADAN database is to provide a centralized repository for peptide binding domain structures and their predicted binding specificities we also tried to make available tools to use these matrices for protein–protein predictions. We note, however, that different domain-types (proline binding, phospho-binding, etc) and different species (single-cell, multi-cellular) will require specific efforts to make the most of the binding predictions provided in ADAN. As an example, for the human SH2 domains, we showed that using phosphorylation data can dramatically improve the prediction of SH2 protein targets since these domains specifically bind phosphorylated tyrosines (Sanchez *et al*., 2008). We also observed that secondary structure filters were more effective in improving the accuracy of SH3 target prediction than SH2 target prediction (Pedro Beltrao, personal communication), suggesting that different domain types might require specific protocols to optimize domain–protein target predictions. For these reason we did not, at this point, attempted to benchmark the accuracy of protein–protein interactions. As we mention below future efforts will be devoted to devise specific protocols to make the best of use of these binding matrices for different domain types and species.

4 DISCUSSION

Most of the structures and functions of globular domains from proteome are yet unknown. However, we can use high-resolution structures from different modular domains in combination with automatic protein design algorithms to predict genome-wide potential interactions of a protein in their physiological context (Fernandez-Ballester and Serrano, 2006; Fernandez-Ballester *et al*., 2009; Kiel *et al*., 2005; Sanchez *et al*., 2008). To facilitate access of the scientific community to this approach, we have created a new database, ADAN, which is composed of a main module that contains all available and continuously updated structural and biochemical information on different modular protein domains, and a second module, Prediadan, that contains all pre-calculated information on protein-protein interactions, as well as some web tools to implement predictions.

Among all common fields and links included in ADAN database (Medline, MINT, PDBsum, Pfam, Protein Data Bank, SMART, Swiss-Prot, etc.), we would like to highlight those of our own elaboration: (i) simplified coordinate files; (ii) pre-calculated scoring matrices, useful for additional predictions; (iii) pre-calculated optimum ligands and models; (iv) pre-calculated list of potential partners in the corresponding genome; (v) interactive web tool to scan a query sequence given by the user, against scoring matrices; and (vi) curated biochemical and thermodynamic data for protein–protein interactions taken from literature.

The ADAN database gathers sequences, isolated structural coordinates, multiple sequence alignments, selection of high resolution ligands for a given interaction domains, pre-computed predictions, etc., in a single resource. This helps in the time consuming and complicated task of modeling a complete set of domains from a genome for the prediction of protein function (Fernandez-Ballester and Serrano, 2006), as already done and validated for SH3 domains from *S.cerevisiae* (Fernandez-Ballester *et al*., 2009).

The most outstanding application of ADAN is the use of the predictive information to guide laboratory experiments. The scoring matrices provided in ADAN can be used to construct optimum peptide ligands and complexes, in combination with FoldX for mutagenesis and theoretical evaluation of energy. Such ligands could be used to rationally target and disrupt a cellular interaction, thus helping to elucidate the role of this interaction in the cell. The predictions can also be used to improve existing interactions or to discover new ones in an easy and reliable process affordable at genome-wide range. The analysis of the information stored in ADAN 'on demand' through the web tool 'Prediction from a query sequence' can also help discover new potential interactions in proteins.

Additionally, Prediadan provides lists of putative partners derived from proteins belonging to the same genome, and sharing (or not) the same subcellular compartment as the query domain. The putative targets were further filtered according to their disorder/globularity conformation and predicted binding energies. These lists, complemented with the already described interaction data from MINT, offer valuable information to validate experimental protein–protein interaction data obtained with high-throughput techniques (phage-display, peptide array, etc.), and have turned into a starting point to plan and conduct experimental investigations of protein function at the genome level.

4.1 Limitations and future work

One of the goals of Prediadan is to offer pre-computed predictions for all modular domains in the database. This is limited by several constrains: (i) our calculations are time-consuming, and several processes cannot be easily automated. Future hardware and software improvements might ease this constraint; (ii) More importantly, we are limited by the amount of structural information available, a problem that can only be solved by an increase in the determination of complex structures or large improvements in homology modeling; (iii) we would like to stress that our FoldX predicted binding energies relate to complex formation in vitro. Future efforts will be directed at developing specific protein interaction protocols for different domain types as well as for different species. We will aim to provide a probabilistic score that integrates the predicted domain–peptide-binding strengths with additional features weighted using traditional machine learning approaches. Some of the features we envision incorporating include: sub-cellular localization, disorder predictions, and experimental interactions data from the MINT database, together with expression/degradation, scaffolding/complex formation and post-translational modifications.

Ideally, we would like to incorporate thermodynamic information for each record in the database in an automated manner. However, the data are dispersed in the literature, and the searching and annotation cannot be automated easily due to the lack of a formal ontology for this kind of data. Even if such ontology was defined and used in the future, an extensive curation effort is still needed to include more existing information from the bibliography in ADAN.

Finally, ADAN will accept coordinate files, so that external users can upload a structure to be analyzed with the predictive tools and generate its own scoring matrix, the list of potential partners, etc. We hope that the analysis of this information generated 'on demand' speeds up the discovery of new protein–protein interactions.

ACKNOWLEDGEMENTS

Special thanks to Dr Ujjwal Das at EBI for linking ADAN to InterPro (http://www.ebi.ac.uk/interpro/). Also thanks to the

'Servicios Informáticos' from the University Miguel Hernandez (Elche, Spain), and to Yann Dublanche from the CRG (Barcelona, Spain) for hosting the ADAN web page. Special thanks to Pilar Aguado-Gimenez for the edition of the manuscript.

Funding: 'Convocatoria de ayudas para proyectos I+D+I para grupos de investigacion emergentes 2007 Generalitat Valenciana' [Exp: GV/2007/025 to G.F.-B.]; 'BANCAJA-UMH' [IP/UR/01 to J.A.E.]; 'Conselleria de Empresa, Universidad y Ciencia de la Generalitat Valenciana' [GV07/017 to J.A.E.]; 'Genome-wide structural and functional analysis of SH3-mediated cellular networks in yeast' [European project: EC 01663 (2005) to L.S]; and 'Functional Proteomics: Towards defining the interaction proteome' [Europen project: EC 505520 to L.S.]. I.E.S. is the recipient of an EMBO Long Term Fellowship. P.B. was supported by a fellowship from the Fundaçao para Ciencia e Tecnologia (SFRH/BDP/41583/2007).

Conflict of Interest: none declared.

REFERENCES

Aloy,P. and Russell,R.B. (2002) Interrogating protein interaction networks through structural biology. *Proc. Natl Acad. Sci. USA*, **99**, 5896–5901.

Apweiler,R. *et al*. (2001) The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res*., **29**, 37–40.

Bairoch,A. *et al*. (2005) The universal protein resource (UniProt). *Nucleic Acids Res*., **33**, D154–D159.

Berman,H. *et al*. (2003) Announcing the worldwide Protein Data Bank. *Nat. Struct. Biol*., **10**, 980.

Ceol,A. *et al*. (2007) DOMINO: a database of domain-peptide interactions. *Nucleic Acids Res*., **35**, D557–D560.

Chatr-aryamontri,A. *et al*. (2007) MINT: the Molecular INTeraction database. *Nucleic Acids Res*., **35**, D572–D574.

Fernandez-Ballester,G. and Serrano,L. (2006) Prediction of protein-protein interaction based on structure. *Methods Mol. Biol*., **340**, 207–234.

Fernandez-Ballester,G. *et al*. (2009) Structure-based prediction of the *Saccharomyces cerevisiae* SH3-ligand interactions. *J. Mol. Biol*., **388**, 902–916.

Fernandez-Ballester,G. *et al*. (2004) The tryptophan switch: changing ligand-binding specificity from type I to type II in SH3 domains. *J. Mol. Biol*., **335**, 619–629.

Finn,R.D. *et al*. (2005) iPfam: visualization of protein-protein interactions in PDB at domain and amino acid resolutions. *Bioinformatics*, **21**, 410–412.

Finn,R.D. *et al*. (2006) Pfam: clans, web tools and services. *Nucleic Acids Res*., **34**, D247–D251.

Gasteiger,E. *et al*. (2001) SWISS-PROT: connecting biomolecular knowledge via a protein database. *Curr. Issues Mol. Biol*., **3**, 47–55.

Gavin,A.C. *et al*. (2006) Proteome survey reveals modularity of the yeast cell machinery. *Nature*, **440**, 631–636.

Gong,S. *et al*. (2005a) A protein domain interaction interface database: InterPare. BMC. *Bioinformatics*, **6**, 207.

Gong,S. *et al*. (2005b) PSIbase: a database of protein structural interactome map (PSIMAP). *Bioinformatics*, **21**, 2541–2543.

Guerois,R. *et al*. (2002) Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J. Mol. Biol*., **320**, 369–387.

Kempkens,O. *et al*. (2006) Computer modelling in combination with in vitro studies reveals similar binding affinities of Drosophila Crumbs for the PDZ domains of Stardust and DmPar-6. Eur. *J. Cell Biol*., **85**, 753–767.

Kiel,C. and Serrano,L. (2006) The ubiquitin domain superfold: structure-based sequence alignments and characterization of binding epitopes. *J. Mol. Biol*., **355**, 821–844.

Kiel,C. *et al*. (2004) A detailed thermodynamic analysis of ras/effector complex interfaces. *J. Mol. Biol*., **340**, 1039–1058.

Kiel,C. *et al*. (2005) Recognizing and defining true Ras binding domains II: in silico prediction based on homology modelling and energy calculations. *J. Mol. Biol*., **348**, 759–75.

Kolsch,V. *et al*. (2007) Control of Drosophila gastrulation by apical localization of adherens junctions and RhoGEF2. *Science*, **315**, 384–386.

Laskowski,R.A. *et al*. (2005) PDBsum more: new summaries and analyses of the known 3D structures of proteins and nucleic acids. *Nucleic Acids Res*., **33**, D266–D268.

Letunic,I. *et al*. (2006) SMART 5: domains in the context of genomes and networks. *Nucleic Acids Res*., **34**, D257–D260.

Linding,R. *et al*. (2003) GlobPlot: exploring protein sequences for globularity and disorder. *Nucleic Acids Res*., **31**, 3701–3708.

Marti-Renom,M.A. *et al*. (2000) Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.*, **29**, 291–325.

Milburn,D. *et al*. (1998) Sequences annotated by structure: a tool to facilitate the use of structural information in sequence analysis. *Protein Eng.*, **11**, 855–859.

Mulder,N.J. *et al*. (2005) InterPro, progress and status in 2005. *Nucleic Acids Res.*, **33**, D201–D205.

Ogmen,U. *et al*. (2005) PRISM: protein interactions by structural matching. *Nucleic Acids Res.*, **33**, W331–W336.

Pieper,U. *et al*. (2006) MODBASE: a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Res*., **34**, D291–D295.

Ponting,C.P. *et al*. (1999) SMART: identification and annotation of domains from signalling and extracellular protein sequences. *Nucleic Acids Res.*, **27**, 229–232.

Prieto,C. and de las Rivas,J. (2006) APID: agile protein interaction DataAnalyzer. *Nucleic Acids Res.*, **34**, W298–W302.

Sanchez,I.E. *et al*. (2008) Genome-wide prediction of SH2 domain targets using structural information and the FoldX algorithm. *PLoS Comput. Biol.*, **4**, e1000052.

Schultz,J. *et al*. (1998) SMART, a simple modular architecture research tool: identification of signaling domains. *Proc. Natl. Acad. Sci. U S A*, **95**, 5857–5864.

Schymkowitz,J. *et al*. (2005) The FoldX web server: an online force field. *Nucleic Acids Res*., **33**, W382–W388.

Tagari,M. *et al*. (2006) E-MSD: improving data deposition and structure quality. *Nucleic Acids Res.*, **34**, D287–D290.

Villanueva,J. *et al*. (2003) Ligand screening by exoproteolysis and mass spectrometry in combination with computer modelling. *J. Mol. Biol.*, **330**, 1039–1048.

Wallace,A.C. *et al*. (1995) LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. *Protein Eng*., **8**, 127–134.