Prediction of Protein-Protein Interaction Based on Structure

Gregorio Fernandez-Ballester[1] & Luis Serrano[2]

[1]Instituto de Biología Molecular y Celular (IBMC). Universidad Miguel Hernandez. Avda. de la Universidad, s/n. Edif. Torregaitán. 03202 Elche (Alicante). Spain.

[2]EMBL, Meyerhofstrasse,1. 69117 Heidelberg, Germany

Abstract

A great challenge in the proteomics and structural genomics era is to predict protein structure and function from sequence, including the identification of biological partners. The development of a procedure to construct position specific scoring matrices for the prediction and identification of sequences with putative significant affinity faces this challenge. The local and web applications used for sequence and structure search, sequence alignment, protein modelling, molecule edition and modification, and scoring matrices construction are described in detail. The methodology is based on the information contained in structural databases, and takes into account the subtle conformational and sequence details that characterize different structures within a family. Using the matrices, the protein sequence databases can be easily scanned to locate putative partners of biological significance. The success of this methodology opens the way for the prediction of protein-protein interaction at genome scale.

Key Words

Bioinformatics, protein-protein interaction, protein modelling, protein prediction, positional scoring matrix, pattern search.

Headings

1. Introduction

A major fraction of the genomes has now been sequenced, and this vast amount of data opens a way for novel methods of analysis of all genes and their products. One of these methods, particularly important, is the prediction and/or characterization of functional interactions between proteins on a genome-wide level. The interest on those parts of proteins ("domains") involved in protein interactions that fold independently of the rest of the molecule it is remarkable, as they have a fully functional interaction activity at high levels, and are commonly crystallised alone or in complex with a polypeptide ("ligands"). However, for most of these domains, the structural and functional features are completely unknown, and their putative roles are only suggested by homology. The information regarding protein-protein interactions and multicomponent systems formation is limited and interaction network maps are incomplete.

Protein-protein interactions are ubiquitous in biology: Transient associations between proteins support a broad range of biological processes, including hormone-receptor binding, protease inhibition, antibody-antigen interaction, signal transduction, correction of misfolding by chaperones, and even enzyme allostery. On the contrary, permanent associations are essential for proteins whose stability or function is defined by multimeric states, as viral capsides, oligomeric enzymes, channel proteins, etc. (*1*)

Protein-protein interactions occur at the surface of a protein and are biophysical phenomena, governed by shape, chemical complementarity, and flexibility of the molecules involved. Assemblies involving proteins that must be independently stable before association, referred as "transient", have interfaces that differ from those in oligomer complexes, referred as "permanent". The permanent association of an oligomer interface tends to be planar, roughly circular in shape, with a high abundance of hydrophobic groups and depletion in charged groups (*2*). On the contrary, transient interfaces more closely resemble the protein

exterior, containing a higher proportion of polar and charged groups, with salt-bridges and hydrogen bonding networks playing a more important role in stabilizing these complexes (*3*).

Nevertheless, both types of interfaces are closely-packed and exhibit a high degree of geometric and electrostatic complementarity. The observations of these interfaces and the apparent consistencies found have led some groups to suggest simple rules for the prediction and location of putative interfaces (*3-7*). However, the properties that make a good interface depend on the type of complex, and should be ranked by different criteria; also, the predictions were more powerful when applied to homodimer than to transient dimers (*2*). The geometric and electrostatic complementarity observed within interfaces has been also the basis of docking studies using proteins of known structure, and putative complexes are refined with electrostatic or chemical criteria to predict the "best" complex.

There are several sequence and structure methods for predicting protein-protein interactions: as examples, i) SPOT, an algorithm to predict ligands (*8*), does not need explicit 3D structures since the interaction database is a multidimensional array containing frequencies at position-specific contacts that explore the probability for a given ligand to bind to a domain; ii) VIP creates virtual interactions profiles (position-specific scoring matrices) from a 3D structure in complex with ligand, and scans sequence databases to seek binding partners of biological relevance (*9*); iii) DOCK is an algorithm that accounts for flexible ligand docking, either with small ligand or protein, followed by virtual screening (*10*).

Analysis of conservation patterns in binding sites benefits from the fact that the residues involved are on the molecular surface and surface conservation is generally low. This potentially high signal-to-noise ratio arises because changes in surface residues do not generally influence folding and overall stability as much as changes in residues at the structural core, so any mutational intolerance that does exist can be detected more easily. Several groups have explored patterns of conservation at binding sites in a systematic way using multiple alignments, and sometimes phylogenetic trees of homologous sequences to map evolutionary information onto datasets of protein structures. If a protein's function is common within a homologous family and essential or advantageous for the survival of the host organism, the maintenance of that function describes the limits to which mutational variation in the sequence may be tolerated. So, if a protein-protein interaction plays an important functional role, it is interesting to study how patterns of evolutionary conservation in the protomer sequences relate to the maintenance of this interaction (*2;11*).

Thermodynamic studies in which the interface is systematically mutated reveal that the distribution of energetically important residues can be uneven across interfaces and concentrated in "hot spots" of binding energy (*12*). There is also evidence that residues distant

from the interface can play a critical role in stabilizing protein-protein interactions. Such residues are believed to be energetically coupled with those directly involved in binding and allow binding energy to propagate through tertiary structure.

In addition to theoretical analyses of crystal structures, a large quantity of experimental studies has provided insights into protein-protein interactions (*3;13-17*). Recently, however, there has been a large increase in the number of known three-dimensional structures that contain protein-protein recognition sites, covering a much broader range of activities than ever, and allowing us to determine the extent of generalization of these rules based on a few structures. The aspects of structure that must be taken into consideration are those related to the stabilization of protein association: The size and chemical character of the protein surface that is buried at interfaces; the packing density of atoms that make contacts across the interface, which expresses complementarity; and polar interactions through hydrogen bonds and interface water molecules. Each of these aspects can be described at the level of the individual atom that forms protein-protein recognition sites, where three classes can be distinguished: Atoms that lose accessibility but do not make direct contacts across the interface; atoms that make direct contacts but remain partly accessible; and atoms that become buried.

For all this, in this chapter we focus on the detailed use and development of bioinformatic applications for the prediction of protein-protein interaction on a structural basis to guess the potential partners of known proteins. The accurate computational measurements of the stability of protein-protein complexes and the improvements of the software for "*in silico*" protein engineering, drug design, mutagenesis, dynamics, etc., allow to tackle the exploration of these surfaces for protein engineering and the prediction of partners, that should help in interpreting experimental data. The methodology focuses on the application of molecular modelling to calculate, manipulate, and predict protein structures and functions. Concepts of structure similarity/overlap, sequence alignment, structure superposition, homology modelling, and molecular docking, which are special concerns of protein biochemists, are considered. Approaches to protein modelling by the use of programs such as Swiss PDB Viewer, and online servers (Swiss Pdb Servers; FoldX) are described. The study is centred in transient domain-ligand interaction, making special emphasis on problems associated with sequence and structure alignments.

The success of these methods provides an invaluable tool to select accurate pools of putative partners for further biochemical/biophysical characterization of proteins. The ultimate goal of these studies is the prediction of protein function at genome scale.

2. Methods

The prediction of protein interactions is a multiple step methodology that requires the comprehension of theoretical concepts, as well as the use of several software applications, either local or from the Internet (*18*). Since all these methodologies could be extended to entire genomes, it is strongly recommended to use scripting languages to automatize the usually tedious and monotonous jobs on each molecule. To launch multiple local jobs we recommend Python (http://www.python.org/) as an effective tool for the automatization of the work. For Internet protein databases search and other purposes, we recommend Perl (http://www.perl.com/) because of its powerfulness and simplicity.

The flow diagram in Figure 1 comprises all steps and methodologies to accomplish our objectives.

2.1. Isolation of domain sequences and domain assignment.

The SMART database at EMBL (http://SMART.embl-heidelberg.de) is a very powerful tool when working with protein-protein interaction domains. The database SMART (*19*) allows the identification and annotation of genetically mobile domains and the analysis of domain architectures, containing more than 400 domain families, extensively annotated with respect to functional class, tertiary structures and functionally important residues. Each domain from non-redundant protein databases is stored in a relational database system with search parameters and taxonomic information. In addition, the web user interface allows searches for proteins containing specific combinations of domains in defined taxa. Useful examples are:

a) Isolation of a given domain from the whole protein sequence:

- Go to SMART web page.
- Paste the protein ID or the sequence of your protein (one letter code) under "Sequence Analysis".
- Press Sequence SMART.

The output gives a picture and a table with the list of functional domains found, the scores and the boundaries. The isolated sequence domains can be accessed individually and stored in text files.

b) Isolation of all sequences of a particular domain found in the whole proteome of a taxon:

- Write the domain ID of your interest in Domain selection under "Architecture analysis".
- Select the taxonomic range in the selection box.

- Press Domain selection.

The output is a list of proteins from a selected organism that contains the domain of interest. Now you can:

- Get all the sequences of the whole protein, or
- Get all the sequences of the isolated domains.
- Save as text file.

The ADAN database (http://adan-embl.ibmc.umh.es/) contains biochemical and structural information on different modular protein domains implied in protein-protein interactions (SH3, SH2, WW, PDZ, PH, methyl transferase, acetyl transferase, WD40, VHS, protein tyrosine phosphatase, PTB, FHA, BRCT, and 14-3-3). The records in the database contain a useful collection of the most common domain features, as well as links to other databases like PDBsum (http://www.ebi.ac.uk/thornton-srv/databases/pdbsum/), Protein Data Bank (http://www.rcsb.org/pdb/) and Swiss-Prot (http://au.expasy.org/). The database also offers the scoring matrices for ligand predictions and putative partners from protein domains when available (see below).

2.2. Homology search

One of the most common analyses done on protein sequences is the similarity/homology search. It allows a mapping of information from known sequences to novel ones, especially the functional sites of the homologous proteins. Dynamic programming has been applied to sequence alignment and related computational problems: A dynamic algorithm finds the best solution by breaking the original problem into smaller sequentially dependent subproblems that are solved first. Each intermediate solution is stored in a table along with a score and the sequence of solutions that yields the highest score is chosen. The search for similarity/homology is well supported by Internet resource tools. A query sequence can be entered to conduct the homology search using BLAST servers at Expasy (http://au.expasy.org/tools/blast/), NCBI (http://www.ncbi.nlm.nih.gov/BLAST/) or EBI (http://www.ebi.ac.uk/blastall/), or WU-BLAST servers at EMBL (http://dove.embl-heidelberg.de/Blast2/) or EBI (http://www.ebi.ac.uk/blast2/).

- Go to web tool.
- Paste or upload the query sequence, normally in FASTA format (sequence of amino acids in one letter code preceded by “>sequence_name”).
- Select the program “blastn” for nucleotides or “blastp” for amino acids.
- Select database “nr” for non-redundant.
- Check other fields/parameters of interest.

- Press Search.

You will receive a submit confirmation, and in a few minutes an output results. The output includes alignment scores (sometimes as a plot), the list of sequences with significant E-values, and the corresponding information (sequences ID, name, score, etc.), and pairwise alignments (if chosen).

The homologue search can be done in most of the BLAST web servers selecting the pdb (Protein Data Bank) as the target database, thus directly obtaining the homologues whose structures are already known. The structures can be downloaded and edited with public software to model the domains of interest.

2.3. Edition of the molecules and template selection

Most comprehensive software programs suitable for protein modelling are commercial packages, some of which are listed in Table 1. The protein modelling is illustrated with freeware programs and online servers, such as Swiss PDB Viewer (http://www.expasy.org/spdbv/), which is an application program that provides a user interface for visualization and analysis of biomolecules in particular proteins (*20*). The program (Swiss PDB viewer or Spdbv) can be freely downloaded, and the user guide is also available. Spdbv implements GROMOS96 force field (*21*) to compute energy and to execute energy minimization by steep descent and conjugate gradient methods.

Typically, the workspace consists of a menu bar, tool icons, and four windows; Main (Display) windows, Control panel, Layers Info and Align window. The Control panel, Layers Info and Align window can be turned on and off from the *Window* menu. The Control panel provides a convenient way to select and manipulate the attributes of individual residues. The first column shows groups included in the structure, including chain name in uppercase letters (A, B, C, etc.), secondary structure recognized by Spdbv, represented by lowercase letters (s, strand; h, alfa-helix), and the names and numbers of all amino acid residues/nucleotides/heteroatoms. The second column and subsequent have check marks that toggle between display/hide the whole group, the side chains, residue labels, dot surfaces (VDW or accessibility), and ribbon (if the check marks are activated). The last column with small square boxes is used to highlight the residue(s) with colours.

The Layers Info Window allows the management of entire layers by turning on and off layer visibility, movement, displayed carbonyl groups, hydrogens, hydrogen bonds, etc. The Align window permits an easy means of manage protein sequences for the homology modelling.

Briefly, the common tools are located at the top of the display window under the menu bar. These include move molecule (translate, zoom, and rotate) tools and general usage (bond distance, bond angle, torsion angle, label, centre, fit, mutation, and torsion) tools. The menus provide a full set of utilities to load, display, select, edit, measure, superpose, modify, etc., the molecules. Additionally, the scripting language provides an invaluable tool to simplify and automatize the repetitive edition or modification of molecules (see http://www.usm.maine.edu/~rhodes/SPVTut/text/DiscuSPV.html to post and/or exchange scripts). This is especially important for studies at genome-wide level.

The selection of templates starts in the Protein Data Bank (http://www.rcsb.org/pdb/) and in fact this is the limiting step for the consecution of a successful prediction on a protein-protein interaction domain. If the sequence homology is low, it is mandatory the existence of several X-Ray, high resolution structures of the domain of interest, either alone or (better) in complex with natural or designed partners, the reason being that small or large conformational rearrangements could take place in the structure of the target sequence that will not be present in the model structure. Thus, the existence of several domain-ligand complexes is of capital importance in the reliability and accuracy of the predictions. Those complexes involving proteins with low sequence homology are more useful since they provide a glimpse of the conformational variability of the protein family. In the absence of this kind of complexes, docking techniques should be used (see below).

The structures of interest can be easily grouped in SMART:

- Go to web tool
- Under "Domains detected by SMART", in "Display domain annotation" fill Domain Name (i.e., SH2, SH3, PDZ, etc.)
- Press Display.
- Select "Structure" and get the list of all structures available for the domain of interest.
- Download the pdb files. Each pdb name in the previous list is directly linked to PDBsum database (http://www.ebi.ac.uk/thornton-srv/databases/pdbsum/) which, in turn, connects with Protein Data Bank.

The structural coordinate files contain sometimes much more structural information than needed for prediction. As an example, SH2 and SH3 domains are usually forming part of tyrosine kinases crystals, including the catalytic domain. Many other times, the coordinates of the domains appear several times in the same pdb file as a consequence of crystal symmetry properties. Other accessory molecules such as heteroatoms can be removed from the structure file if they are not involved in the protein interaction. These "cleaned" structures can be

directly downloaded from the ADAN database, or can be easily isolated with Spdbv following these steps:

- Load the molecule in Swiss PDB viewer: *File* / *Open PDB file*.
- Select the residues comprising the domain of interest, as well as the natural ligand, if present. Use Control Panel or Alignment window. Activate first these windows in Window/Control Panel or Window/Alignment.
- Save the selected part of the molecule. *File* / *Save* / *Save selected residues*.

It is necessary to stress the importance of keeping the structures of the natural or designed ligands in complex with the domain of interest, when present in the crystal.

2.4. Clustering of templates

At this point we have a set of structures representative of the domain of interest that can be compared, grouped and classified according to different structural and/or sequence features. This important step allows the connection between proteins with known and unknown structures through homology modelling. A failed clustering gives unrealistic models and unsuccessful predictions even when the target sequences and templates are quite homologous. Unfortunately, there is no automatic way to carry out these steps, and it should be made by hand after careful observation of structural motives, key positions repetition, etc. Thus, a previous study of the templates is required, including either multiple superposition or multiple sequence alignment of the structures and their sequences.

Superposition can be easily made with Spdbv following these steps:

- Open the template1 molecule and colour the whole molecule if desired.
- Open the template2 molecule.
- Make active the template2 layer by clicking its name in Layer Info, or Alignment, or Control Panel windows (three ways to do the same).
- Invoke the *Iterative Magic Fit* tool from the *Fit* menu.
- Choose the Auto Fit Options (α-carbon or backbone), and press OK. The template2 molecule is superimposed on template2 molecule.
- Check the root mean square distance (rmsd < 1.5Å is better), and press *Fit* / *Improve Fit* to improve the superimposition (lowering rms distance).
- Get the correct alignment by clicking *Fit* / *Generate Structural Alignment*. The movement of the cursor on the residues in Alignment window causes the residue in the Display window to blink to orange colour, allowing the easy visualization of the overlapping molecules.

- Save each of the overlapped molecules separately (*File* / *Save Layer*) as individual structure or save the superimposed molecules together (*File* / *Save Project*) as overlapped structures.

Interesting examples of previous studies are provided by the SH3 domain. The characteristics of SH3 sequences taken into account are: i) Key positions that determine the protein folding, and later on the function, thus making the sequence belong to a given domain. In SH3, nine core positions, two very well conserved Gly and three binding positions (*22;23*) can be distinguished. These positions must be conserved (hydrophobic residues in the core) or identical (Glys and Trp, Pro and Phe in the binding pocket), and should be checked manually after sequence alignment (Figure 2A); ii) The length of the loops *RT* and *n-Src* involved in binding. Since this feature can be involved in ligand recognition, it is important the clustering of templates by loop length. Figure 2B illustrates the groups obtained for SH3 structures classified according to *n-Src* length, since the *RT* loop is quite constant. Other insertions and deletions in the SH3 sequence not involved in binding (i.e. *distal* loop) are not taken into consideration.

Relevant structural characteristics of SH3 domains are: i) the Trp switch, where a very well conserved Trp in the binding pocket of SH3 can tilt a few degrees depending of the kind of residues in the immediate vicinity and the nature of the ligand (*24*). The consequence of this small Trp movement governs in part the binding specificity of SH3 to poly-Pro ligands type I or type II. So, depending on the orientation of this Trp in our template, we will be able to predict binding to type I or type II, but not both with the same structure (Figure 3); ii) Motive YXY or YXF in the *RT*-loop of SH3 domain. The second aromatic in this motive is part of the binding pocket of SH3 and is pointing to other important residues in the pocket. Some templates having Phe in the motive cannot accommodate Tyr (after homology modelling, see below) because either the extra hydroxyl group could clash with the conserved Trp, or the hydroxyl becomes uncompensated into the pocket, or both. On the contrary, the templates having Tyr in the motive do not have steric clashes after substitution by Phe, but residues previously forming hydrogen bonds with Tyr could become uncompensated into the binding pocket.

As a consequence of both sequence and structural features, the construction of chimeras from structural templates would be necessary to fulfil all the requirements needed for modelling a sequence target. A good example is shown in Figure 1C, were the yeast protein rvs167 SH3 domain is modelled with a chimera formed with 1SHF.PDB and 1OOT.PDB templates. The former (1SHF.PDB) fits very well with protein yeast in the *RT* loop (motives YDY and DL), but presented a deletion in *n-Src* and a insertion in *distal* loops. The later

(1OOT.PDB) fits well in *n-Src* and *distal* loop, but fails in the motive YDY to YSF (see boxes in the Figure 2C).

The construction of chimeras can be easily accomplished with Spdbv as follows:

- Load the two templates selected to generate the chimera. Load first the template contributing with the N-terminal fragment.
- Superimpose following the fully automatic method (see above) or, if necessary, the manual one, to get a good overlapping of the connection point between the two templates to build the chimera.
- For manual superimposition select suitable parts of the molecules, and take care of selecting the same number of residues in both layers (the number of residues selected can be followed in the last column of Layers Info window).
- Select *Fit Molecules (from selection)* in the *Fit* menu. Select *Auto Fit Options* (backbone) and press OK.
- Align sequences with *Fit* / *Generate structural alignment*.
- In the Align window select the residues in the first molecule contributing with the N-terminal. Select the residues of the second molecule contributing with the C-terminal. They should superimpose well to avoid peptidic bond distortion.
- Assemble the two selected fragments with *Create Merged Layer from Selection* in the *Edit* menu. An extra layer ("_merge_") appears with the chimera.
- Activate the merged layer and select all residues in *Select* /*All*.
- Renumber and/or rename the chain with *Edit* / *Rename Current Layer*.
- Save the merged layer with *File* / *Save Layer*.

2.5. Selection of ligands

The importance of the availability of high resolution domain-ligand complexes for prediction of protein interaction is already mentioned. These structures help to understand in detail the hydrophobic and hydrophilic interaction map and confirm the key residues important for binding (*6*). At the same time these structures validates the surface prediction studies (*5;7*), and allows the accurate prediction of potential partners based on structure.

The templates containing ligands are filtered according to the resolution quality, trying to recruit ligands with 2.5Å resolution or better. These ligands are "cleaned" to remove parts of the protein not interacting with the domain, and comprising no more than 7 to 10 residues long for extended ligands. The isolated ligands are usually grouped into categories, depending of the nature and family features, and stored for later use (see below).

As an example, ligands binding to SH3 are grouped into type I, type II, type I' (*24*) or other types (*25*) up to 24 different ligands.

2.6. Sequence alignment

This step allows the assignments of the target sequences to the clusters previously obtained for the structures, so that each sequence is only linked to the templates included in the cluster, but not to all templates. The alignment of the target sequence inside its group permits a more accurate and realistic alignment, which is complemented with the chimera construction when there are no good candidates as templates.

The pairwise or multiple sequence alignment can be accomplished with ClustalW, the most commonly used program (Table 2):

- Go to web tool to get the query form.
- Paste or upload the sequences in FASTA format.
- Press Send.

The output includes pairwise alignment scores, multiple sequence alignment and tree file. Check carefully if the new alignment fits all the requirements of your domain (see SH3 examples).

2.7. Homology modelling

Protein modelling aims to predict the 3D structures of proteins from their amino acid sequences, using related sequences for which structures are available. The prediction of protein structures is based on two complementary approaches that can be used in conjunction: i) Knowledge-based model combining sequence data to structure information, such as homology modelling (*26;27*). The methodology modifies closely related functionally analogous sequence molecules (orthologous) whose 3D structure has been previously elucidated, and a putative 3D structure (model) of a protein from a known 3D structure is obtained. Thus, functionally analogous proteins with homologous sequences will have closely related structures with identical tertiary folding patterns. ii) Energy-based calculations through theoretical models and energy minimization, such as *ab initio* prediction (*28*). Energy-based structure prediction relies on energy minimization and molecular dynamics. The method is faced with the problem of a large number of possible multiple minima, making the traversal of the conformational space difficult, and making the detection of the real energy minimum or native conformation uncertain. Residues are changed in the sequence with minimal disturbance to the geometry, and energy minimization optimizes the altered structure.

Although the success of homology modelling is satisfactory when sequence homology is greater than 50% , the structure homology may remain significant even if sequence homology is low: 3D structures are better conserved than the residue sequence. The region between 20-30% sequence identities is less certain, since part of the problem of homology modelling is the correct alignment of unknown and target proteins. Care should be taken with the important key residues within a structurally conserved common fold.

A methodology to explore sequence identities below 25% (remote homologues) is threading techniques (*29*), where a sequence of unknown structure is threaded into a sequence of known structure, and the fitness of the sequences for that structure is assessed.

Basically, the homology modelling consists of four steps:

1. Start from the known sequences.

2. Assemble the new sequence onto the template backbone from different, known homologous structures.

3. Optimize the structures.

4. Select the structures with better stability energy.

The modelling of a protein structure from its sequence against the known 3D structure of homologous protein(s) in the homology modelling can be attempted with Spdbv as follows:

- Prepare your previously obtained alignment between the target sequence and the homologue pdb structure. You will use it later.
- Save your target sequence in FASTA format in a text file (i.e. “target.txt”). Add extra sequence if you want to model the ligand in complex. In this case your template should have the ligand coordinates.
- Choose the *Load Raw Sequence to Model* tool from the *SwissModel* menu and open “target.txt”. The target sequence appears on the Display window as a long helix.
- Open the template structure of the reference molecule (template.pdb).
- Select the *Fit Raw Sequence* tool from the *Fit* menu. The α-helix changes to the structure overlapping the reference structure. Centre the molecule in the Display window.
- Click the little arrow beside the question mark of the Alignment window to view a plot of threading energies. Click *smooth* to set smooth to 1 and check *SwissModel / Update threading display now* tool. Select *Color / By Threading Energy* to display threading energy profile of the structure (the mean force potential energy of the polypeptide chain increases with colour varying from blue to red).

- The alignment of the target sequence onto the template can be manually refined on the Alignment window by translating residues or inserting and removing gaps.
- Select a residue in the Alignment window and use the space bar to insert a gap or the backspace key to remove a gap. Select a group of residues and use left/right arrow keys to displace a gap. Do the same for the template if necessary. A long bond appears in the Display window.
- Displace the ligand sequence in the target in order to align with the ligand sequence in the template. The ligand sequence can be identical or different to the template since it will be explored later.
- Break the connection between domain and ligand in the target. Use menu *Build*, then *Break backbone* and pick one of the backbone atoms connecting domain and ligand target.
- Activate the target structure and change chain name (i.e. A for domain and B for ligand) and numbering, if desired.
- Perform two operations: *Select / Select_aa Making Clashes*, and *Tools / Fix Selected Side* Chains / Quick and Dirty. Repeat the process to decrease the number of amino acid residues making clashes.
- Make sure that Swiss Model settings, under *Preferences* menu has the correct server information: Modelling server (http://swissmodel.expasy.org/cgi-bin/sm-submit-request.cgi); template server (http://swissmodel.expasy.org/cgi-bin/blastexpdb.cgi), your name and your e-mail correctly written (very important).
- Submit your model to SwissPdbServer using Optimise (project) mode for optimization. Go to *Submit Modelling Request* under *SwissModel* menu. You are asked to give a name for your model (i.e. my_project.htm in html format). An additional project file is automatically saved as proj_my_project.pdb as coordinates.
- Your Internet browser opens and shows the webpage to be submitted. Fill the information regarding your Swiss Model Project and check the appropriate Result options (usually Swiss PDB viewer Mode and What-if check) and press Send Request. You will get a notification of successful submission.
- You will receive in your e-mail several messages. One of them has your model in complex with the desired ligand. The two structures (domain and ligand) are

already merged if you have decided to model both together. Open the received project and save the first layer to isolate your final model.

The models at this step require visual inspection and evaluation in terms of energy (see below) to check the quality of the models. We should look for inconsistencies in the core (i.e. polar groups, strong clashes between core residues), in the binding pocket (i.e. incompatible residues, bad rotamers), or in the loop movements among different templates (i.e. clashes, obstacle for ligand binding), etc.

2.8. Evaluation of the models in terms of energy

The use of fast and reliable protein force field is an efficient tool to evaluate the delicate balance between the different energy terms that contribute to protein stability (*30;31*). Many different force fields are constructed for predicting protein stability changes, ranging from force fields based on pure statistical analysis of structural sequence preferences (*32;33*), or force fields based on multiple sequence alignments (*23;34*), to detailed molecular dynamics force fields (*35;36*). These force fields can be divided into three major categories: i) Those using a physical effective energy function (PEEF), ii) Those based on statistical potentials for which energies are derived from the frequency of residue or atom contacts in the protein database (SEEF), and iii) Those using empirical data obtained from experiments on proteins (EEEF).

There are several molecular modelling packages that implement force fields that can be used for the evaluation of the models in terms of energy: Spdbv implements Gromos96 force field; InsightII (Accelrys, Inc.) implements CVFF, CFF91 and AMBER force fields; Sybyl (Tripos, Inc.) implements MM2, amberall40 and amberuni40 force fields, etc.

In addition, FoldX (*37;38*), a computer 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 weighted 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.

The FoldX energy function includes terms that have been found to be important for protein stability. The free energy of unfolding (ΔG) of a target protein is calculated using equation (1):

$$\Delta G = \Delta G_{vdw} + \Delta G_{solvH} + \Delta G_{solvP} + \Delta G_{wb} + \Delta G_{hbond} + \Delta G_{el} + \Delta G_{kon} + \cdot T\Delta S_{mc} + \cdot T\Delta S_{sc} + T\Delta S_{tr} \quad (1)$$

where ΔG_{vdw} is the sum of the Van der Waals contributions of all atoms with respect to the same interactions with the solvent. ΔG_{solvH} and ΔG_{solvP} are the difference in solvation energy

for apolar and polar groups respectively when going from the unfolded to the folded state. ΔG_{hbond} is the free energy difference between the formation of an intra-molecular hydrogen-bond compared to inter-molecular hydrogen-bond formation (with solvent). ΔG_{wb} is the extra stabilising free energy provided by a water molecule making more than one hydrogen-bond to the protein (water bridges) that cannot be taken into account with non-explicit solvent approximations. ΔG_{el} is the electrostatic contribution of charged groups, including the helix dipole. ΔG_{kon} reflects the effect of electrostatic interactions on *k*on (39) term for protein complexes. ΔS_{mc} is the entropy cost for fixing the backbone in the folded state. This term is dependent on the intrinsic tendency of a particular amino acid to adopt certain dihedral angles. ΔS_{sc} is the entropic cost of fixing a side chain in a particular conformation. Finally, ΔS_{tr} is the loss of translational and rotational entropy upon making the complex.

The FoldX web server (http://foldx.embl.de) gives the user the stability energy, the interaction network of the different energy components per interaction type and per residue, and/or the energy contribution per chain in a protein-protein complex.

The careful examination of these energy values permits the selection of high quality molecular models and opens the way for successful prediction.

2.9. Ligand superposition

When a large amount of structures in complex with ligands are available, it is quite easy to add these ligands to models to perform predictions: Just follow the protocol used for generating chimeras to create the complexes (see above). It should be noted that the superposition has to be accurate and realistic, and has to be adapted to the actual target. For SH3 domains as an example, manual superposition is used by selecting 9 residues in the core and four residues in the binding pocket, but this can be quite different for other domains, such as PDZ or SH2.

When complexes are not available, it is necessary to use docking techniques, thus assuming a great problem inherent to the methodology: The peptidic ligand conformation. Molecular docking aims to fit two interacting molecules by exploring the binding modes of their topographic or energy-based features consideration that lead to favorable interactions. Ligand-receptor interaction is an important initial step in protein prediction and function. The structure of ligand-receptor complex profoundly affects the specificity and efficiency of protein action. The molecular docking performs the computational prediction of the ligand-receptor interaction and the structures of ligand-receptor complexes, usually computing the van der Waals and the Coulombic energy contributions between all atoms of the two molecules.

There are two classes of strategies for docking a ligand to a receptor. The first class uses a whole ligand molecule as a starting point and employs a search algorithm to explore the energy profile of the ligand at the binding site, looking for optimal solutions for a specific scoring function. The search algorithms include geometric complementary match, simulated annealing, molecular dynamics, and genetic algorithms. Representative examples are DOCK3.5 (*40*), AutoDock (*41*), and GOLD (*42*). The second class starts by placing one or several fragments (substructures) of a ligand into a binding pocket, and then, it constructs the rest of the molecule in the site. Representative examples are DOCK4.0 (*43*), FlexX (*44*), LUDI (*45*), and GROWMOL (*46*). Table 3 shows some Internet resources for molecule docking.

The use of docking methods from isolated structures is, however, very difficult and risky since the conformations of the isolated ligand and the receptor can be different (or very different) after complex formation. For this reason, it is strongly recommended the use of ligand-receptor structure complexes of high resolution. The methodological use of the docking methods is out of the scope of this chapter.

2.10. Selection of complexes

Finally, we have a set of models in complex with ligands that should be evaluated. First of all, the positions in the ligand should be mutated to Ala in order to minimize the clashes within the complex, but also to normalize all the positions in the same ligand, and among different ligands.

Mutations can be easily accomplished with Swiss PDB viewer following these steps in the active molecule:

- Press the mutation icon
- Pick the amino acid to be mutated. The list of amino acids is displayed.
- Select the desired amino acid. A rotamer is selected automatically.
- Change the rotamer by clicking the small black arrows under the mutation icon.
- Press the mutation icon again and select OK.

Refine the structure by removing clashes:

- Press *Select / aa making clashes*.
- Fix clashing residues with *Tools / Fix selected Sidechains* and *Quick and Dirty* or *Exhaustive Search* (no more than 10 residues).
- Save your mutated structure (*File / Save / Layer*).

The complexes must be explored, looking for incompatibilities between domains and ligands. Mainly, strong clashes between ligand and domain backbones should be avoided, and

in this case, the complexes should be discarded. Regarding side chains, small clashes can be tolerated for external residues that are able to adopt different rotamers, but cannot for residues important for binding.

Again, the energy calculation with FoldX provides a great tool to explore the complexes in terms of energy. The stability energy of the complex (E_s) and the binding energy (E_b) between domain and ligand can be calculated in FoldX web server. The E_b can be calculated under AnalyseComplex mode as:

$$E_b = E_s - \sum (E_{sA} + E_{sB})$$

where E_{sA} and E_{sB} refer to the stability energy of isolated chains A (domain) and B (ligand), respectively.

Finally, we have a set of high quality complexes, selected for ligand and domain compatibility, and prepared to predict the optimal ligand, construct the scoring matrices and search in the databases.

2.11. Modelling from secondary and tertiary structure predictions

The modelling of proteins that have homologues in the Protein Data Bank is very easy. However, not all proteins have homologues of known structure. This case requires an additional secondary and/or tertiary prediction to build models and guess the putative partners. The flow diagram in Figure 1 shows an alternative route when the sequences of interest have no homologue structures. First of all, a prediction of secondary structure is required, an old technique not exempt of many problems (the early methods suffered from a lack of data, predictions were performed on single sequences rather than families of homologous sequences, and there were relatively few known 3D structures to derive parameters). Nowadays, this technique is more accurate and aims to determine the probable placement of secondary structural elements along the sequence. The prediction is made at several levels: i) Secondary structure prediction, expecting three-states (helix, strand, rest) with an accuracy ranging 72-76% for water-soluble globular proteins; ii) Solvent accessibility prediction; iii) Transmembrane helix prediction, expecting overall two-states (transmembrane, non-transmembrane) with an accuracy higher than 95%; and iv) Globularity, that identifies inter-domain segments containing linear motifs and apparently ordered regions that do not contain any recognised domain. Most common servers for protein secondary structure are depicted in Table 4.

Even with no homologue of known 3D structure, it may be possible to find a suitable fold for you protein among known 3D structures by folding recognition methods (Table 5). There are many approaches, but the unifying theme is to try and find folds that are compatible

with a particular sequence. These methods combine 1D (or even 3D) sequence profiles coupled with secondary structure and contact capacity potential information to thread a protein sequence through the set of structures and predict the fold (*47*).

The alignments of sequence onto tertiary structure from fold recognition methods may be inaccurate. After the identification of a remote homologue, it is convenient to edit the alignment around variable regions and consider the alignment of secondary structures. Check that: i) The residues predicted to be buried or exposed are aligned with those known to be buried/exposed in the template structure; ii) Hydrogen bonds networks are not disrupted in beta-sheet structures; and iii) The residues properties (i.e. size, polarity, hydrophobicity) are conserved as much as possible in the alignment.

If the fold can not be yet recognized, there are methods to try *ab initio* structure predictions or at least predictions that do not rely on a template. *Ab initio* prediction of protein 3D structures is not "possible" at present, and a general solution to the protein folding problem is not likely to be found in the near future. However, some methods have been developed to try the prediction of the structure of proteins starting from the sequence by calculating: i) Secondary structure in the form of three states (helix, extended, loop); ii) Local conformation (backbone torsion angles phi and psi); iii) Supersecondary structure for strands an beta-turns; and iv) Tertiary structure, in the form of coordinates. These methods (Table 6) are based on hidden Markov model for local and secondary structure prediction, based on the I-sites Library.

Finally, we have models that can be evaluated in terms of energy, then following the route in the prediction diagram (Figure 1).

2.12. Scoring matrices construction

The models in complex with poly-Ala ligands are now ready to use for scoring matrices construction (Figure 4). Basically, each position in the ligand is explored by systematic mutation of the Ala to the 20 natural amino acids and further energy evaluation of the resulting structures (stabilization and binding energies). The energies obtained are correlated with the ability of a residue (in a ligand, in a position) to improve the ligand-domain interaction. Several assumptions have to be made:

a) Every position in the ligand is treated and computed separately within the ligand. This simplification is of great importance to save hard disk space and computational time, and is based on the fact that most ligands binds to the domain in an "extended" conformation (i.e. SH2, PDZ, poly-Pro helix in SH3, etc), that makes contiguous residues to point to different directions and to interact with different set of residues in the domain. In the case in which

there is energy coupling between two positions in the ligand, other more sophisticated approaches like mean field, dead-end elimination, branch and bound or Montecarlo techniques could be used to explore many-fold sequence combinations.

b) The domain residues in the immediate vicinity of the mutating ligand position should be allowed to change their rotamer to accommodate the new environment after mutation. This is also important to avoid strong van der Waals clashes and to optimize hydrophobic/hydrophilic interactions. The most important positions in the binding pocket probably will not change their rotamers (i.e. W, P and Y in the SH3 pocket), but other positions involved in binding could be adapted to improve the interaction.

Methodologically, there is not a direct way to build these structures easy an automatically with commercial applications. The construction of 20 structures per position, with a mean of 9 positions per ligand, and with the overall use of 10 ligands, gives a total of 1800 structure files. This number is big enough to dissuade anybody to build these structures by hand. Most modelling packages include scripting capabilities that can be adapted to:

1. Mutate a position in a ligand to the first of 20 amino acids.
2. Relax the surrounded domain positions.
3. Mutate the same position to the first amino acid again. This is made to avoid conformational traps.
4. Relax again the surrounded domain positions.
5. Save the coordinates of the new structure.
6. Repeat 1 to 5 for every position in a ligand.
7. Repeat 1 to 6 for all ligands.
8. Repeat 1 to 7 for all complexes.
9. Evaluate the stabilization and binding energies for all structures with FoldX.

At this point we have a set of structures that represents a complete screening of the ligand-domain interactions. The evaluation of the binding energy of these structures provides the link between ligand position, type of residue and binding improvement. This quantification results in scoring matrices (see Figure 5).

The matrices are corrected by adding internal van der Waals clashes of the interface residues with their own chains to the binding energy and normalised to the lower value (becoming 0). The lower the energy value, the better the ligand-domain interaction. Once having the matrices, we model the best ligand by taking the most favourable amino acid at each position (now all positions of the ligand at the same time) and evaluating its binding energy. This will be used as a reference for a particular matrix. Note in the example (Figure 5) that some positions are more tolerant than others to accommodate different residues. This is a

reflection of the role of the positions in the binding interaction, being more permissive the positions that point to the solvent than those pointing to the hydrophobic pocket.

2.13. Database search and hits filtering

The scoring matrices provide the link between computational predictions of partners and the localization of these resulting motives in the genome databases, so that we can use these data to scan de genome, looking for the sequence/s that better fit in the modelled SH3 domains, thus guessing function. The simplest way to do this is the generation of Prosite patterns and the use of the web application in Prosite web page (http://au.expasy.org/tools/scanprosite/). Prosite patterns should follow a specific format (see Prosite *format* link in the webpage), and can be easily derived from the scoring matrices. As an example, this pattern is obtained from the scoring matrix of Figure 5:

Pattern: **[PIL]-X-[HKRM]-X-[PHQ]-P-[HDPWFRS]-[MPLKR]-[PW]**

The ScanProsite tool allows to look protein sequence(s) for the occurrence of patterns, profiles and rules stored in the Prosite database, but also to search protein database(s) for hits by specific motif(s) (*48*). The last feature is the one used for our purposes to scan genomes. To search a genome with a pattern:

- GO to web application.
- Enter one pattern under *PROSITE pattern(s)/profile(s) to scan for*.
- Select the database to search: Swiss-Prot and TrEMBL.
- Write your taxon in *Taxonomic lineage (OC) / species (OS)*.
- Press *STAR THE SCAN*.

After genome scanning, the results are presented in the screen. The information included contains protein name, SwissProt code numbers, protein length, position and sequence of the hits, etc., as well as the link to get the whole protein. Depending on the patterns entered, the results obtained can vary from a few of proteins to several hundreds. The results should be filtered because probably most of the hits obtained are not real. There is not a fixed rule to do this filtering step. What follows is just an example that can be modified depending of the domain, the taxon, the available biochemical data, etc.

The simplest way to filter the hits is follows:

- Get the sequences of the proteins found by ScanProsite.
- Send the sequences to a protein secondary structure prediction server.
- Send the sequences to predict globularity/disorder (http://globplot.embl.de/). GlobPlot is a web service that allows the user to plot the tendency within the query protein for order/globularity and disorder.

- Match secondary structure and order/disorder predictions (per protein).

Now, the hit sequence is localized in the match and checked for coincidences: The hit must be located in a region corresponding to both a disordered secondary structure prediction and to an unstructured part of the protein (Beltrao & Serrano, submitted for publication). Whether a hit missing one residue in these regions is selected or not depends on each case. Again, there are no fixed rules and the limits can be established and modified according to the necessities.

The remaining hits can be filtered again with more criteria: The most important is probably the experimental data, when available. All panning experiments to search putative specific ligands of the domains are suitable to validate the prediction. As an example, peptide libraries displayed in filamentous phages or large-scale two-hybrid interaction tests are used for these purposes (*49*). There are also available, for some domains (i.e. SH3), different computational prediction data that serve for comparisons and mutual validation (*8;9*). A further filtering criterion can be the data regarding the subcellular localization of the proteins in a taxon. It is possible that a predicted hit that fulfils all the conditions to interact with the domain, is confined to a different compartment, so that they never meet in the cell. In this case, the hit should be discarded.

The scoring matrices can be also used to evaluate the probability of interaction between a domain and a given peptide, by summing the positional free energy of each amino acid in the sequence and comparing with the optimum ligand.

3. Conclusion

As a general thought, the prediction of protein-protein interactions based on structure should be viewed as a new tool to guess protein function, to improve database annotation, and to design rational experiments to understand protein network interactions.

Acknowledgements

The authors thanks Dr. Pilar Aguado Giménez for editing the manuscript.

References

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

2. Valdar, W. S. and Thornton, J. M. (2001) Protein-protein interfaces: analysis of amino acid conservation in homodimers, *Proteins 42*, 108-124.

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

4. Jones, S. and Thornton, J. M. (1997) Prediction of protein-protein interaction sites using patch analysis, *J. Mol. Biol. 272*, 133-143.

5. Aloy, P., Querol, E., Aviles, F. X., and Sternberg, M. J. (2001) Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking, *J. Mol. Biol. 311*, 395-408.

6. Stein, A., Russell, R. B., and Aloy, P. (2005) 3did: interacting protein domains of known three-dimensional structure, *Nucleic Acids Res. 33 Database Issue*, D413-D417.

7. Aloy, P., Bottcher, B., Ceulemans, H., Leutwein, C., Mellwig, C., Fischer, S., Gavin, A. C., Bork, P., Superti-Furga, G., Serrano, L., and Russell, R. B. (2004) Structure-based assembly of protein complexes in yeast, *Science 303*, 2026-2029.

8. Brannetti, B., Via, A., Cestra, G., Cesareni, G., and Helmer-Citterich, M. (2000) SH3-SPOT: an algorithm to predict preferred ligands to different members of the SH3 gene family, *J. Mol. Biol. 298*, 313-328.

9. Wollacott, A. M. and Desjarlais, J. R. (2001) Virtual interaction profiles of proteins, *J. Mol. Biol. 313*, 317-342.

10. Meng, E. C., Gschwend, D. A., Blaney, J. M., and Kuntz, I. D. (1993) Orientational sampling and rigid-body minimization in molecular docking, *Proteins 17*, 266-278.

11. Lichtarge, O. and Sowa, M. E. (2002) Evolutionary predictions of binding surfaces and interactions, *Curr. Opin. Struct. Biol. 12*, 21-27.

12. Bogan, A. A. and Thorn, K. S. (1998) Anatomy of hot spots in protein interfaces, *J. Mol. Biol. 280*, 1-9.

13. Schreiber, G. and Fersht, A. R. (1995) Energetics of protein-protein interactions: analysis of the barnase-barstar interface by single mutations and double mutant cycles, *J. Mol. Biol. 248*, 478-486.

14. Janin, J. and Chothia, C. (1990) The structure of protein-protein recognition sites, *J. Biol. Chem. 265*, 16027-16030.

15. Janin, J. (1995) Principles of protein-protein recognition from structure to thermodynamics, *Biochimie 77*, 497-505.

16. Jones, S. and Thornton, J. M. (1996) Principles of protein-protein interactions, *Proc. Natl. Acad. Sci. U. S. A 93*, 13-20.

17. Tsai, C. J., Lin, S. L., Wolfson, H. J., and Nussinov, R. (1996) A dataset of protein-protein interfaces generated with a sequence-order-independent comparison technique, *J. Mol. Biol. 260*, 604-620.

18. Tsai, C. S. (2002) Molecular modelling: protein modelling, in *An introduction to computational biochemistry* (Han, L., Ed.) pp 315-342, John Wiley & Sons, Inc., New York.

19. Letunic, I., Copley, R. R., Schmidt, S., Ciccarelli, F. D., Doerks, T., Schultz, J., Ponting, C. P., and Bork, P. (2004) SMART 4.0: towards genomic data integration, *Nucleic Acids Res. 32*, D142-D144.

20. Guex, N. and Peitsch, M. C. (1997) SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling, *Electrophoresis 18*, 2714-2723.

21. van Gunsteren, W. F. and Mark, A. E. (1992) Prediction of the activity and stability effects of site-directed mutagenesis on a protein core, *J. Mol. Biol. 227*, 389-395.

22. Northey, J. G., Di Nardo, A. A., and Davidson, A. R. (2002) Hydrophobic core packing in the SH3 domain folding transition state, *Nat. Struct. Biol. 9*, 126-130.

23. Larson, S. M. and Davidson, A. R. (2000) The identification of conserved interactions within the SH3 domain by alignment of sequences and structures, *Protein Sci. 9*, 2170-2180.

24. Fernandez-Ballester, G., Blanes-Mira, C., and Serrano, L. (2004) The tryptophan switch: changing ligand-binding specificity from type I to type II in SH3 domains, *J. Mol. Biol. 335*, 619-629.

25. Cesareni, G., Panni, S., Nardelli, G., and Castagnoli, L. (2002) Can we infer peptide recognition specificity mediated by SH3 domains?, *FEBS Lett. 513*, 38-44.

26. Hilbert, M., Bohm, G., and Jaenicke, R. (1993) Structural relationships of homologous proteins as a fundamental principle in homology modeling, *Proteins 17*, 138-151.

27. Chinea, G., Padron, G., Hooft, R. W., Sander, C., and Vriend, G. (1995) The use of position-specific rotamers in model building by homology, *Proteins 23*, 415-421.

28. Bonneau, R. and Baker, D. (2001) Ab initio protein structure prediction: progress and prospects, *Annu. Rev. Biophys. Biomol. Struct. 30*, 173-189.

29. Bryant, S. H. and Altschul, S. F. (1995) Statistics of sequence-structure threading, *Curr. Opin. Struct. Biol. 5*, 236-244.

30. Murphy, K. P. and Freire, E. (1992) Thermodynamics of structural stability and cooperative folding behavior in proteins, *Adv. Protein Chem. 43*, 313-361.

31. Pace, C. N., Shirley, B. A., McNutt, M., and Gajiwala, K. (1996) Forces contributing to the conformational stability of proteins, *FASEB J. 10*, 75-83.

32. Sippl, M. J. (1995) Knowledge-based potentials for proteins, *Curr. Opin. Struct. Biol. 5*, 229-235.

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

34. Bordo, D. and Argos, P. (1991) Suggestions for "safe" residue substitutions in site-directed mutagenesis, *J. Mol. Biol. 217*, 721-729.

35. Prevost, M., Wodak, S. J., Tidor, B., and Karplus, M. (1991) Contribution of the hydrophobic effect to protein stability: analysis based on simulations of the Ile-96----Ala mutation in barnase, *Proc. Natl. Acad. Sci. U. S. A 88*, 10880-10884.

36. Pitera, J. W. and Kollman, P. A. (2000) Exhaustive mutagenesis in silico: multicoordinate free energy calculations on proteins and peptides, *Proteins 41*, 385-397.

37. Guerois, R., Nielsen, J. E., and Serrano, L. (2002) Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations, *J. Mol. Biol. 320*, 369-387.

38. Kiel, C., Serrano, L., and Herrmann, C. (2004) A detailed thermodynamic analysis of ras/effector complex interfaces, *J. Mol. Biol. 340*, 1039-1058.

39. Vijayakumar, M., Wong, K. Y., Schreiber, G., Fersht, A. R., Szabo, A., and Zhou, H. X. (1998) Electrostatic enhancement of diffusion-controlled protein-protein association: comparison of theory and experiment on barnase and barstar, *J. Mol. Biol. 278*, 1015-1024.

40. Kuntz, I. D., Blaney, J. M., Oatley, S. J., Langridge, R., and Ferrin, T. E. (1982) A geometric approach to macromolecule-ligand interactions, *J. Mol. Biol. 161*, 269-288.

41. Morris, G. M., Goodsell, D. S., Huey, R., and Olson, A. J. (1996) Distributed automated docking of flexible ligands to proteins: parallel applications of AutoDock 2.4, *J. Comput. Aided Mol. Des 10*, 293-304.

42. Jones, G., Willett, P., Glen, R. C., Leach, A. R., and Taylor, R. (1997) Development and validation of a genetic algorithm for flexible docking, *J. Mol. Biol. 267*, 727-748.

43. Ewing, T. J., Makino, S., Skillman, A. G., and Kuntz, I. D. (2001) DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases, *J. Comput. Aided Mol. Des 15*, 411-428.

44. Rarey, M., Kramer, B., Lengauer, T., and Klebe, G. (1996) A fast flexible docking method using an incremental construction algorithm, *J. Mol. Biol. 261*, 470-489.

45. Bohm, H. J. (1992) The computer program LUDI: a new method for the de novo design of enzyme inhibitors, *J. Comput. Aided Mol. Des 6*, 61-78.

46. Bohacek, R. S. and McMartin, C. (1997) Modern computational chemistry and drug discovery: structure generating programs, *Curr. Opin. Chem. Biol. 1*, 157-161.

47. Jones, D. and Thornton, J. (1993) Protein fold recognition, *J. Comput. Aided Mol. Des 7*, 439-456.

48. Gattiker, A., Gasteiger, E., and Bairoch, A. (2002) ScanProsite: a reference implementation of a PROSITE scanning tool, *Appl. Bioinformatics. 1*, 107-108.

49. Tong, A. H., Drees, B., Nardelli, G., Bader, G. D., Brannetti, B., Castagnoli, L., Evangelista, M., Ferracuti, S., Nelson, B., Paoluzi, S., Quondam, M., Zucconi, A., Hogue, C. W., Fields, S., Boone, C., and Cesareni, G. (2002) A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules, *Science 295*, 321-324.

Figures

Figure 1. Flow chart of methodologies for protein prediction based on structure. (1). Isolation of domain sequences and domain assignment. (2) Homology search. (3) Edition of the molecules and template selection. (4) Clustering of templates. (5) Selection of ligands. (6) Sequence alignment. (7) Homology modelling. (8) Evaluation of the models in terms of energy. (9) Ligand superposition. (10) Selection of complexes. (11) Modelling from secondary and tertiary structure predictions. (12) Scoring matrices construction. (13). Database search and hits filtering.

Figure 2. Sequence characteristics tested for clustering of SH3 templates. Panel A shows the key positions in SH3 core (*), Gly (%), binding ($) and motive YXY (&). Panel B shows different *n-Src* loop lengths in different SH3 templates. Panel C shows the criteria used for chimera construction from two SH3 templates, 1SHF.pdb and 1OOT.pdb. The resulting chimera was optimal to model the yeast SH3 rvs167 protein.

Figure 3. Structural characteristics tested for clustering of SH3 templates. The figure shows the different conserved Trp orientation observed upon binding of ligand type I to Abl SH3 domain (template 1ABO.PDB, black arrows) or ligand type II to C-Crk N-terminal SH3 domain (template 1CKA.PDB).

Figure 4. The picture represents a yeast SH3 domain in complex with a type I ligand 10 residues long. All 20 natural amino acids are placed in each position, and the surrounded residues are allowed to relax. The positions in the ligand are processed independently, and the coordinate files are generated and evaluated in terms of energy.

Figure 5. Scoring matrix from a yeast SH3 domain. Rows represent position in the ligand, and columns represent the 20 natural amino acids. The matrices are built by generating a set of structures containing systematic mutations in the ligand and further evaluation of the binding energy. The values are usually corrected and normalised to the lower one (the lower, the better).

Tables

Table 1. Some protein modelling commercial and non commercial packages.

Modelling Package	URL Address
Insight II (Accelrys, Inc.)	http://www.accelrys.com/sim/
Chem3D (CambridgeSoft Corp)	http://www.camsoft.com
HyperChem (Hypercube, Inc.)	http://www.hyper.com
SYBYL (Tripos, Inc.)	http://www.tripos.com
SPDBV (GSK)	http://www.expasy.org/spdbv/
Wavefunction, Inc.	http://www.wavefun.com
MOE (CCG, Inc.)	http://www.chemcomp.com
Modeller (UCSF)	http://salilab.org/modeller/modeller.html
WHAT IF (CMBI)	http://swift.cmbi.ru.nl/whatif/

Table 2. Clustalw servers for sequence alignment. Some servers for structure alignments are also included.

CLUSTALW Servers	URL Address
CLUSTALW (EBI)	http://www.ebi.ac.uk/clustalw/
CLUSTALW (PBIL)	http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_clustalw.html
MUSCA	http://cbcsrv.watson.ibm.com/Tmsa.html
T-Coffee	http://www.ch.embnet.org/software/TCoffee.html
Dialign	http://bibiserv.techfak.uni-bielefeld.de/dialign/
Structure alignments	**URL Address**
CE / CL	http://cl.sdsc.edu
Dali	http://www.ebi.ac.uk/dali/
TMAP	http://www.mbb.ki.se/tmap/
MAMMOTH	http://ub.cbm.uam.es/mammoth/mult/index.php

Table 3. Available resources for docking purposes.

Docking applications	URL Address
GRAMM	http://www.bioinformatics.ku.edu/vakser/gramm/
HEX	http://www.csd.abdn.ac.uk/hex/
FlexX	http://www.biosolveit.de/software/
DOCK	http://dock.compbio.ucsf.edu/
AUTODOCK	http://www.scripps.edu/mb/olson/doc/autodock/
LIGIN	http://swift.cmbi.kun.nl/swift/ligin/
ICM-Docking	http://www.molsoft.com/docking.html
3D-Dock	http://www.bmm.icnet.uk/docking/
ZDOCK, RDOCK	http://zlab.bu.edu/zdock/index.shtml
Bielefeld Software	http://www.techfak.uni-bielefeld.de/~sneumann/agaiprot/
Molfit	http://www.weizmann.ac.il/Chemical_Research_Support//molfit/
3D-JIGSAW	http://www.bmm.icnet.uk/~3djigsaw/

Table 4. Secondary structure prediction servers and links.

Secondary structure prediction	URL Address
GOR	http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_gor4.html
HNN	http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_nn.html
DSC	http://www.aber.ac.uk/~phiwww/prof/
PredictProtein	http://www.embl-heidelberg.de/predictprotein/predictprotein.html
nnPredict	http://www.cmpharm.ucsf.edu/~nomi/nnpredict.html
PSA	http://bmerc-www.bu.edu/psa/
BCM-PSSP	http://dot.imgen.bcm.tmc.edu/psspredict ion/pssp.html
JPred	http://www.compbio.dundee.ac.uk/~www-jpred/submit.html
Predator	http://bioweb.pasteur.fr/seqanal/interfaces/predator-simple.html

Table 5. Folding recognition servers and links.

Folding recognition	URL Address
3D-PSSM	http://www.sbg.bio.ic.ac.uk/~3dpssm/
UCLA-DOE	http://fold.doe-mbi.ucla.edu/
LOOPP	http://cbsu.tc.cornell.edu/software/loopp/
PROSPECT Pro!	http://www.bioinformaticssolutions.com/products/prospect.php
123D	http://123d.ncifcrf.gov/123D+.html
UCSC HMM	http://www.cse.ucsc.edu/research/compbio/HMM-apps/
FFAS03	http://ffas.ljcrf.edu/ffas-cgi/cgi/ffas.pl
PSPC	http://globin.bio.warwick.ac.uk/%7Ejones/threader.html

Table 6. *Ab initio* structure prediction sites.

Ab initio prediction	URL Address
Rosetta/HMMSTR	http://www.bioinfo.rpi.edu/%7Ebystrc/hmmstr/server.php
I-Sites	http://www.brunel.ac.uk/depts/bio/project/biocomp/mak_fan/isites.htm

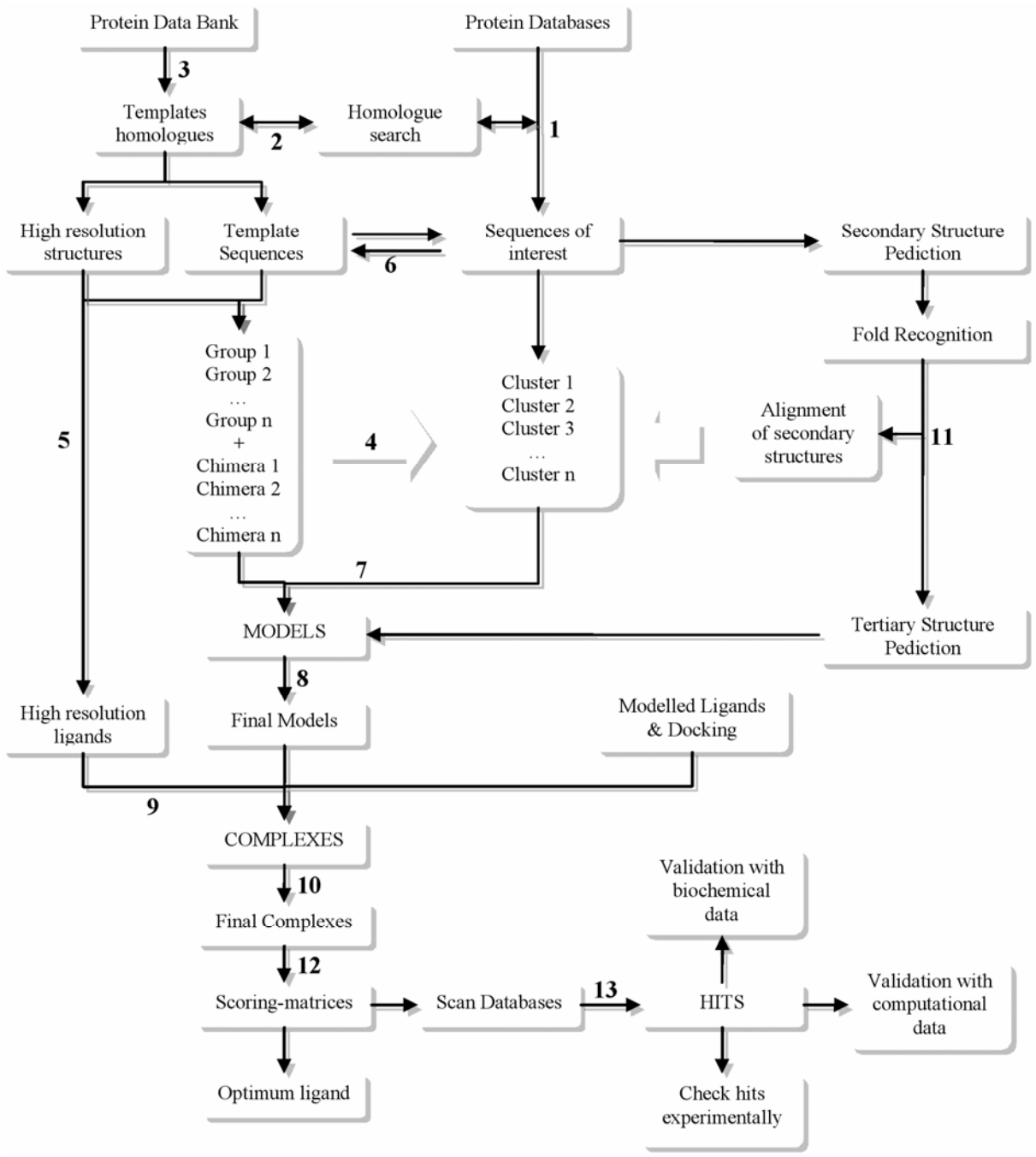
Protein Data Bank
Protein Databases
3
Templates homologues
2
Homologue search
1
High resolution structures
Template Sequences
6
Sequences of interest
Secondary Structure Pediction
Fold Recognition
Group 1
Group 2
…
Group n
+
Chimera 1
Chimera 2
…
Chimera n
5
4
Cluster 1
Cluster 2
Cluster 3
…
Cluster n
Alignment of secondary structures
11
7
MODELS
Tertiary Structure Pediction
8
High resolution ligands
Final Models
Modelled Ligands & Docking
9
COMPLEXES
10
Final Complexes
12
Scoring-matrices
Scan Databases
13
HITS
Validation with biochemical data
Validation with computational data
Optimum ligand
Check hits experimentally

FIGURE 1

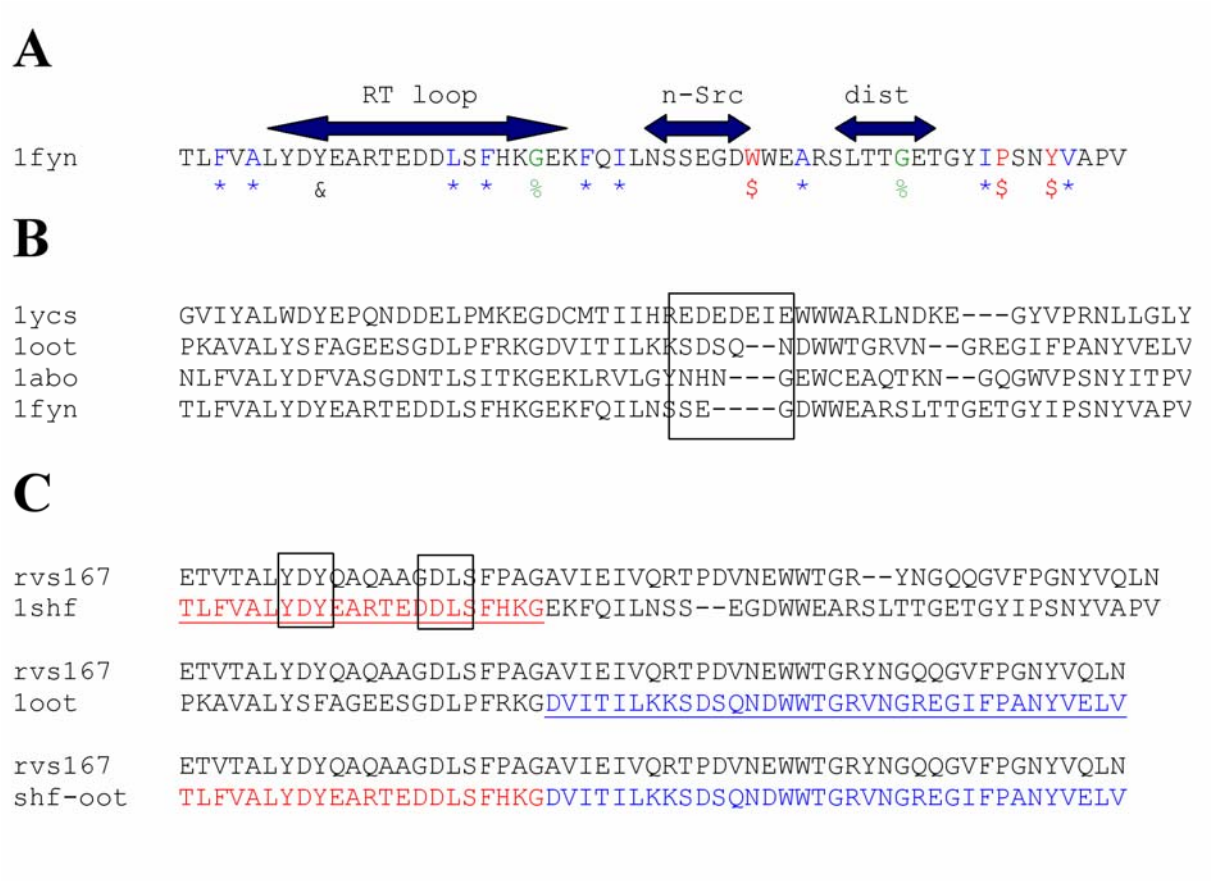
A
RT loop
n-Src
dist
1fyn TLFVALYDYEARTEDDLSFHKGEKFQILNSSEGDWWEARSLTTGETGYIPSNYVAPV
B
1ycs GVIYALWDYEPQNDDELPMKEGDCMTIIHREDEDEIEWWWARLNDKE---GYVPRNLLGLY
1oot PKAVALYSFAGEESGDLPFRKGDVITILKKSDSQ--NDWWTGRVN--GREGIFPANYVELV
1abo NLFVALYDFVASGDNTLSITKGEKLRVLGYNHN---GEWCEAQTKN--GQGWVPSNYITPV
1fyn TLFVALYDYEARTEDDLSFHKGEKFQILNSSE----GDWWEARSLTTGETGYIPSNYVAPV
C
rvs167 ETVTALYDYQAQAAGDLSFPAGAVIEIVQRTPDVNEWWTGR--YNGQQGVFPGNYVQLN
1shf TLFVALYDYEARTEDDLSFHKGEKFQILNSS--EGDWWEARSLTTGETGYIPSNYVAPV
rvs167 ETVTALYDYQAQAAGDLSFPAGAVIEIVQRTPDVNEWWTGRYNGQQGVFPGNYVQLN
1oot PKAVALYSFAGEESGDLPFRKGDVITILKKSDSQNDWWTGRVNGREGIFPANYVELV
rvs167 ETVTALYDYQAQAAGDLSFPAGAVIEIVQRTPDVNEWWTGRYNGQQGVFPGNYVQLN
shf-oot TLFVALYDYEARTEDDLSFHKGDVITILKKSDSQNDWWTGRVNGREGIFPANYVELV

FIGURE 2

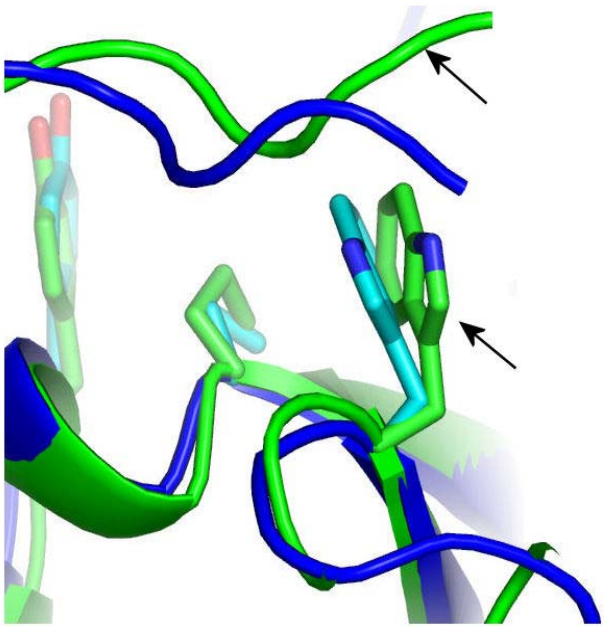

FIGURE 3

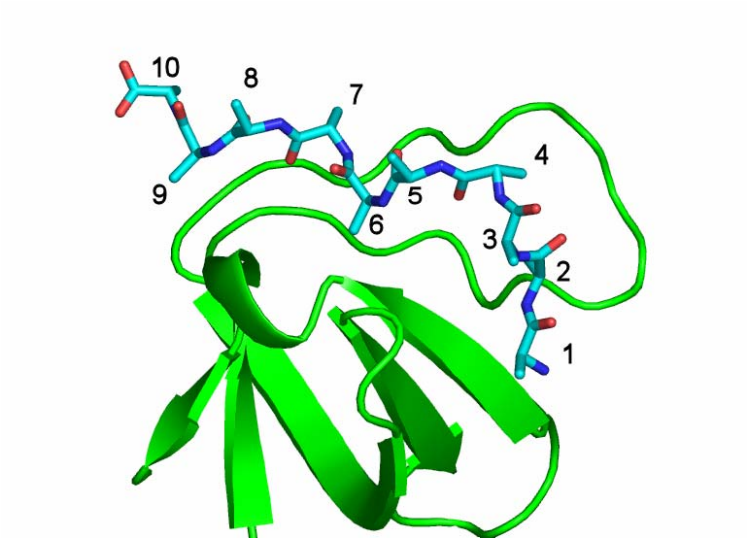

FIGURE 4

#	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL	optimum	
1	1.60	2.18	1.22	2.59	1.16	1.92	2.51	1.58	0.00	2.16	3.24	0.60	1.51	1.23	1.69	2.26	2.26	2.60	1.34	1.80	H	
2	1.21	0.79	1.31	2.62	0.70	1.47	1.10	1.48	0.52	0.22	0.27	1.44	1.21	2.45	0.00	1.97	2.22	0.87	2.75	0.80	P	IL
3	0.52	0.18	0.56	1.00	0.71	0.47	0.64	0.66	0.43	0.48	0.45	0.13	0.49	0.46	0.00	0.52	0.59	0.58	0.51	0.57	P	KRHLFQIM
4	1.15	0.38	0.87	0.70	1.06	0.81	1.09	0.77	0.00	0.54	0.78	0.29	0.37	0.91	3.25	0.71	0.99	0.98	0.97	0.95	H	KMR
5	0.66	0.23	0.30	0.34	0.83	0.23	0.36	0.19	0.32	0.41	0.34	0.10	0.18	0.27	0.38	0.58	0.73	0.00	0.36	0.65	W	KMGQRFNHDLYEPI
6	1.46	4.0E+005	1.59	3.33	0.99	0.38	1.41	2.83	0.18	17.34	5.0E+003	4.61	114.13	1.12	0.00	1.90	11.09	4.0E+005	2.79	13.03	P	HQ
7	0.91	0.76	0.67	1.15	1.16	1.03	1.19	1.22	0.63	0.92	0.85	0.70	0.84	0.63	0.00	0.84	1.46	0.76	0.62	1.18	P	
8	0.61	0.47	0.77	0.20	0.78	0.60	0.78	0.65	0.00	0.68	0.65	0.57	0.59	0.47	0.29	0.48	1.04	0.46	0.66	0.94	H	DPWFRS
9	1.20	0.41	1.48	1.55	1.02	0.54	1.07	1.45	85.87	0.51	0.31	0.39	0.00	5.0E+004	0.24	1.71	1.75	9.8E+004	2.0E+007	0.60	M	PLKR
10	1.53	0.58	1.23	1.85	1.19	1.58	1.64	1.59	0.98	1.30	1.10	0.86	0.87	0.73	0.00	1.63	1.82	0.40	1.20	1.75	P	W

FIGURE 5