Prediction of Ras-effector interactions using position energy matrices

Christina Kiel* and Luis Serrano
EMBL-CRG Systems Biology Unit, CRG-Centre de Regulacio Genomica, Dr Aiguader 88, 08003 Barcelona, Spain

Received and revised on May 11, 2007; accepted on June 16, 2007
Advance Access publication June 28, 2007
Associate Editor: Anna Tramontano

ABSTRACT

Motivation: One of the more challenging problems in biology is to determine the cellular protein interaction network. Progress has been made to predict protein–protein interactions based on structural information, assuming that structural similar proteins interact in a similar way. In a previous publication, we have determined a genome-wide Ras-effector interaction network based on homology models, with a high accuracy of predicting binding and non-binding domains. However, for a prediction on a genome-wide scale, homology modelling is a time-consuming process. Therefore, we here successfully developed a faster method using position energy matrices, where based on different Ras-effector X-ray template structures, all amino acids in the effector binding domain are sequentially mutated to all other amino acid residues and the effect on binding energy is calculated. Those pre-calculated matrices can then be used to score for binding any Ras or effector sequences.

Results: Based on position energy matrices, the sequences of putative Ras-binding domains can be scanned quickly to calculate an energy sum value. By calibrating energy sum values using quantitative experimental binding data, thresholds can be defined and thus non-binding domains can be excluded quickly. Sequences which have energy sum values above this threshold are considered to be potential binding domains, and could be further analysed using homology modelling. This prediction method could be applied to other protein families sharing conserved interaction types, in order to determine in a fast way large scale cellular protein interaction networks. Thus, it could have an important impact on future in silico structural genomics approaches, in particular with regard to increasing structural proteomics efforts, aiming to determine all possible domain folds and interaction types.

Availability: All matrices are deposited in the ADAN database (http://adan-embl.ibmc.umh.es/).
Contact: christina.kiel@crg.es
Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

Prediction of protein–protein interactions based on structural information is an important tool in systems biology (Aloy and Russell, 2006; Beltrao et al., 2007). The prediction method is based on the finding that structural similar proteins usually interact in a similar way (Aloy and Russell, 2005; Aloy et al., 2005). One way of structure-based prediction is to generate homology models of proteins which have a similar sequence, and calculate the interaction energy of the modelled complex using protein design algorithms (Kiel et al., 2005). In order to generate a homology model, the amino acid side chains of a protein complex are replaced by the corresponding amino acid side chains of two other proteins which belong to the same families, but where no structural information is available. Prediction of protein interactions using homology modelling and energy calculations has been successfully done for protein family members sharing sequence homology (Kiel et al., 2005; Kiel et al., 2007), with high levels of accuracy (Kiel et al., 2007).

Recently we have determined a genome-wide Ras-effector interaction network, for 20 Ras-like proteins in complex with 50 putative Ras-binding domains, based on homology modelling and energy calculations (Kiel et al., 2007). Predicting protein interactions based on homology modelling although accurate (Kiel et al., 2005; Kiel et al., 2007) is a computer time-consuming method, taking ~20 min per model, since it requires full side chain reconstruction of the complex. This makes in silico structural genomics a time-consuming task. Here, we develop a faster method which implies the calculation of position energy matrices, and we applied this method to the prediction of Ras-effector interactions for which complete models have been made (Kiel et al., 2007). These complexes are an example of a conserved interaction type: the interface of Ras proteins in complex with the Ubiquitin-like domain of effector proteins is mainly formed by two β-sheets (β2 and β3 of Ras and β1 and β2 of the effector protein), and by the first helix of the Ubiquitin-like domain.

Ras proteins belong to the superfamily of small GTP-binding proteins where more than 150 proteins have been identified up to now (Colicelli et al., 2004; Takai et al., 2001; Vetter and Wittinghofer, 2001). Members of the Ras-subfamily play an important role in various signal transduction pathways, like proliferation and differentiation. Similar to all other guanine nucleotide binding proteins, Ras proteins have the ability to cycle between an inactive GDP- and an active GTP-bound form (Bourne et al., 1990; Bourne et al., 1991). In the active form the interaction with effectors is possible, which
Prediction of Ras-effector interactions

2 METHODS AND RESULTS

2.1 Generation of 120 template structures

A general flow scheme for the generation of template structures and matrices is shown in Figure 1.

All Ras-binding domains of effector proteins which have been structurally characterized up to now show a similar topology, the ubiquitin-like fold and complex formation between Ras and effector proteins is mainly mediated by the interaction of two β-sheets (β2 and β3 of Ras and β1 and β2 of the effector protein), and by the first helix of the effector proteins (Bunney et al., 2006; Huang et al., 1998; Nassar et al., 1995, 1996; Pacold et al., 2000; Scheffzek et al., 2001; Vetter et al., 1999) (Supplementary Fig. 1a). Although the general binding mode is similar there are small changes in the details of how the interface is formed. Thus, in order to account for this conformational flexibility, all available template structures are used. Out of the seven available X-ray Ras-effector complex structures, we have selected five [which are of good quality (<3Å resolution) and are complete] to be modified to be used as template structures (similar as done in Kiel et al., 2007). From the Ras-RalGDS structure (PDB entry: 1LFD) (Huang et al., 1998), we selected both complexes in the crystallographic unit, molecules AB (template T1a) and molecules CD (template T1b). The Ras-Pi3K complex (PDB entry: 1HE8) (Pacold et al., 2000) was selected as template T2, the Ras-spByr structure (PDB entry: 1K8R) (Scheffzek et al., 2001) as template T3, the Rap-RafRBD complex (PDB entry: IGUA) (Nassar et al., 1996) as template T4 and the Ras-PLCeRA2 complex (PDB entry: 2C5L) (Bunney et al., 2006) as template T5.

Since the main contribution to complex affinity in Ras-effector complexes originates from residues in the first three secondary structure elements of the effector domain (β1, β2 and α1), the six selected template structures (T1a, T1b, T2, T3, T4, T5) have been further modified by deleting all secondary structure elements and connecting loops except β1, β2, and α1 (Kiel et al., 2007). An overlay of all template Ras-effector structures used in this study is shown in Supplementary Figure 1b. The lengths of all template structures are summarized in Supplementary Table 1.

For comparison purposes we have selected the same 20 Ras-like proteins as in the previous study (Kiel et al., 2007). In order to generate the Ras-like proteins, we have only mutated residues which are in the interface or at the edge of the interface (Supplementary Fig. 2a) using version 2.7 of FoldX. The rest of the protein was kept unchanged. The reason been that all Ras members so far are very similar in sequence and structure [in our previous work (Kiel et al., 2007) we checked that the interface residues of other Ras members were compatible with the rest of the WT Ras structure]. All indicated positions were mutated according to the alignment of the selected Ras family members as shown in Supplementary Figure 2b. Thus we have generated 120 modified template structures, which are then used for generating the position scan energy matrices.

2.2 Energy matrices

The 120 templates structures are then further modified by mutating all residues in the Ras-binding domain to alanine. The reason for this is that RBD domains are quite different in sequence and therefore there could be incompatibility of certain amino acids within a certain sequence context. On each of the 120 Ras-alanine effectors all residues at the interface of the effector are mutated independently to all other 20 amino acid residues and the energy difference in binding calculated using FoldX (Guerois et al., 2002; Schymkowitz et al., 2005a, Schymkowitz et al., 2005b). An example of the output for a position energy matrix is shown in Supplementary Figure 3. Depending on the position and the type of amino acid in the Ras-binding domain, the side chain residues contribute either favourable (positive values) or non-favourable (negative values) to complex formation. All matrices are deposited in the ADAN database (http://adan-embl.ibmc.umbc.es/).
Experimental domain will bind a particular Ras protein, we used the experimental thresholds for each of the six template structures to decide if an effector values between the six different template structures. In order to define distribution of energy sum values after selecting the best energy sum effector template for a particular sequence. In Figure 2, we show the experimental binding information for complexes of non-template sequences, and in blue we show experimental binding information for complexes of template sequences (e.g. RalGDS, Raf, PLCeRA2 and PI3K). The size of the squares correlates with the magnitude of interaction.

Using this information a threshold of \(-16.59\) kcal/mol was found for predicting non-binding sequences, if one false positive is excluded (Fig. 2). Sequences below this threshold are predicted to be non-binding domains and can be excluded quickly. Sequences which have energy sum values above this threshold are considered to be potential binding domains, and could be further analysed using homology modelling. Energy sum values above \(-22\) kcal/mol have a high probability to be binding domains (81%). Thus, we take this threshold to test how good our prediction success is to predict binding domains.

### 2.5 Accuracy of the prediction method and saved time compared to homology models

The accuracy of predicting non-binding Ras-binding domains was calculated by using a set of 70 pull-down experimental binding information from literature and our previous publication (Kiel et al., 2007). For 12 complexes where non-binding was predicted based on non-template sequences, and where pull-down results are available, 10 complexes showed non-binding in pull-down experiments or gel-bands with less than 2-fold intensity compared to the control, and in only 2 cases binding was found experimentally. Thus, the accuracy of excluding non-binding is good (0.67), not counting the cases with less than 2-fold intensity (Fig. 3). In comparison, the accuracy of predicting non-binding domains using homology modelling was 0.90 (Kiel et al., 2007), where for 30 non-binding predicted complexes only two showed binding in experiments (Fig. 3). Using the binding threshold for energy sum values of \(-22\), a prediction accuracy of predicting binding domains was found to be good as well (0.73), but more non-binding complexes are found here, where binding was predicted experimentally, and thus the accuracy is better using homology modelling (Fig. 3).

In our previous study, we have predicted the interaction of 50 potential Ras-binding domains in complex with 20 different Ras proteins using homology modelling and energy calculations (Kiel et al., 2007). Out of these 1000 interactions, 409 domains are predicted not to bind (including the 120 template structures we used in this study to generate the matrices (Kiel et al., 2007)). Using energy matrices, 309 complexes (from 1000) are below the threshold of \(-16.59\) (non-binding). Total 174 of these are predicted to be non-binding using homology modelling, 130 are predicted to be in the twilight zone and only 3 are predicted to bind in homology modelling (Kiel et al., 2007). It is important to mention that many of the energy values of homology models in the twilight zone are close to the threshold of non-binding, and that also one third of the homology models predicted to be in the twilight zone did non-bind, as shown by pull-down experiments.

### 3 DISCUSSION

We have used position energy matrices in order to predict Ras-effector interactions. Based on energy sum values and calibration with experimental binding information, potential Ras-binding domains can be scanned quickly and non-binding domains can be sorted. Sequences which have energy sum values above this threshold are considered to be potential binding domains, and could be further analysed using homology modelling. The accuracy of predicting binding and fluorescence-based methods were used. The experimental information was added into the diagrams where the binding energies of all Ras and effector combinations tested on a particular template were displayed (Fig. 2). In red we show experimental binding information for complexes of non-template sequences, and in blue we show experimental binding information for complexes of template sequences (e.g. RalGDS, Raf, PLCeRA2 and PI3K). The size of the squares correlates with the magnitude of interaction.
New methods as found for the Ubiquitin-like domain of the PlexinB1 receptor SMART/Pfam databases, due to very low sequence homology, recorded in one of the five Ubiquitin-subfamilies in the with a Ubiquitin-like topology might exist, but which are not further sequences, which are predicted to have a Ubiquitin-like fold, and predict whether they could interact with one of the 20 different Ras proteins. In our previous study, 50 sequences belonging to one of the five Ubiquitin-like fold subfamilies, mainly sequences of the RA, RBD and PI3K_rbs subfamilies. However, there are many more sequences with a predicted Ubiquitin-like topology. Some of them we have excluded them so far, since the sequence cannot be modelled reliable with one of the available template structures (Kiel et al., 2007). But with new template X-ray Ras-effector complex structures available, more sequences could be modelled. Further it is important to mention, that many more domains with a Ubiquitin-like topology might exist, but which are not recorded in one of the five Ubiquitin-subfamilies in the SMART/Pfam databases, due to very low sequence homology, as found for the Ubiquitin-like domain of the PlexinB1 receptor (Kiel and Serrano, 2006; Tong and Buck, 2005). New methods in domain predictions, structure-based alignments, secondary structure prediction tools, etc. (Kiel and Serrano, 2006), might reveal new Ubiquitin-like sequences, which could be screened for binding with the available matrices.

Prediction of protein–protein interactions based on energy matrices could be very important in future in silico approaches in structural genomics and systems biology, which aim to fully predict, model, and understand the protein interaction network of a cell. After the complete sequencing of several genomes, the completion of 3D structures of all possible domain folds and interaction types would open the possibility to fully model and understand biological systems (Aloy and Russell, 2004; Banci et al., 2007; Bork and Serrano, 2005). Structural proteomics efforts have already found 700–800 different folds (Aloy and Russell, 2004) of the predicted 1000 domain folds in nature (Chotia, 1992) and ~2000 of the predicted 10000 domain–domain interaction types (Aloy and Russell, 2004). Matrices of all interaction types can be stored and sequences belonging to a similar domain family can be scanned quickly, in order to filter out clear non-binding domains.

Using matrices will probably be even more important as a tool in predicting protein–peptide interactions in silico. The interface of protein–peptide complexes is usually more flexible within a similar domain family, and therefore more template structures are needed to model all sequences (Fernandez-Ballester and Serrano, 2006). Thus, if many matrices of protein–peptide databases are already stored in a database, potential binding peptides can be scanned quickly. In addition, large scale screening for possible target peptides does not involve any domain prediction, alignments or loop modelling, and therefore, whole genomes can be scanned quickly (Fernandez-Ballester et al., in preparation).

### ACKNOWLEDGEMENT

We thank the EU for financial support (INTERACTION PROTEOME, grant-No. LSHG-CT-2003-505520).

**Conflict of Interest:** none declared.

### REFERENCES


Tong, Y. and Buck, M. (2005) 1H, 15N and 13C Resonance assignments and secondary structure determination reveal that the minimal Rac1 GTPase binding domain of plexin-B1 has a ubiquitin fold. J. Biomol. NMR, 31, 369–370.