Phospho3D: a database of three-dimensional structures of protein phosphorylation sites

Andreas Zanzoni*, Gabriele Ausiello, Allegra Via, Pier Federico Gherardini and Manuela Helmer-Citterich

Centre for Molecular Bioinformatics, Department of Biology, University of Rome ‘Tor Vergata’, Rome 00133, Italy

Received August 11, 2006; Revised October 13, 2006; Accepted October 17, 2006

ABSTRACT
Phosphorylation is a key regulatory mechanism in biology. It involves the modification of specific protein residues (serine, threonine, tyrosine) by protein kinases (PKs), one of the largest protein families in eukaryotes (2). The amount of data produced by screening assays is growing continuously, making the development of computational tools for collecting and analysing experimental data a pivotal task in the study of phosphorylation. Here we present Phospho3D, a database of 3D structures of phosphorylation sites, which stores information retrieved from the phospho.ELM database and is enriched with structural information and annotations at the residue level. The database also collects the results of a large-scale structural comparison procedure providing clues for the identification of new putative phosphorylation sites.

INTRODUCTION
The phosphorylation of specific protein residues is a crucial event in the regulation of several cellular processes, operating on activation, deactivation or recognition of the target protein. A great deal of eukaryotic proteins (~30% in the human genome) undergo this reversible post-translational modification (1). Phosphorylation on serine/threonine or tyrosine residues is accomplished by protein kinases (PKs), one of the largest protein families, comprising 1.5–2.5% of all eukaryotic genes (2).

Although the amount of data produced in various screening assays is steadily growing (3–6), experimental identification of phosphorylation sites remains a difficult and time-consuming task. Hence, the implementation of computational tools proves to be very useful for collecting and analysing experimental data.

Several sequence-based methods to predict phosphorylation sites were developed using different computational approaches such as regular expressions with context-based rules (7), position-specific scoring matrices (PSSMs) (8), artificial neural networks (9,10), support vector machines (SVMs) (11,12), hidden Markov models (13) and iterative statistical methods (14). All these methods are based on the hypothesis that the sequence surrounding the phosphorylated residue represents the main determinant for kinase specificity. They are reasonably accurate and work well with a number of specific kinases. However, the specificity determinants and rules remain elusive for a large number of protein kinases that display a number of substrates sharing little or no sequence similarity in the known phosphopeptides. We propose that, at least in some cases, the rules of kinase specificity may reside in the presence of structural determinants which only occasionally overlap with sequence consensi and which might be independent of the residue order in protein sequences.

Here we describe Phospho3D, a database of 3D structures of phosphorylation sites. It collects information retrieved from the phospho.ELM database (15) and is enriched with structural information and annotations at the residue level. In addition, the database stores the results of a large-scale local structural comparison which suggest functional annotation of phosphorylation sites by 3D similarity. Cases of significant structural similarity between phosphorylation sites may indicate that they are phosphorylated by the same kinase.

DATABASE CONSTRUCTION AND CONTENT
The Phospho3D database was constructed by collecting data from the phospho.ELM database which gathers experimentally verified phosphorylation sites manually extracted from the literature. The phospho.ELM dataset used in this work (version 4.0) contains 5314 phosphorylation sites, or instances, belonging to 1805 different sequences.

The correspondence between phospho.ELM sequences and the Protein Data Bank (PDB) chains was established via the Seq2Struct resource (16), an exhaustive collection of...
annotated links between SwissProt-TrEMBL and PDB sequences. Links are based on sequence alignment using pre-established highly reliable thresholds. From a list of 4530 sequence–structure links (for further details see website documentation), only the ones having the phosphorylable residue in the alignment region were retained, this resulting in 2726 instances (166 unique phospho.ELM instances on 1219 protein chains).

The basic information stored in Phospho3D consists of the instance, its flanking sequence (10 residues) and any residue whose distance from the instance does not exceed 12 Å thus defining a 3D neighbourhood which we define as zone.

For each zone, annotation at the residue level is provided, namely solvent accessibility supplied by the NACCESS program (17), secondary structure assignment given by the DSSP program (18) and residue conservation as from the Consurf-HSSP database (19).

Users can also retrieve information extracted from the phospho.ELM dataset; for instance, the Medline reference PMID and, when available, the kinase(s) that phosphorylate(s) the given site.

In addition, for each zone the results of a large-scale local structural comparison versus a representative dataset of PDB (20) protein chains from eukaryotic organisms are also given. The comparison was carried out using the Query3D sequence/fold independent algorithm (21). Structural matches are assessed by two criteria: structural similarity and biochemical similarity. The structural similarity demands that matching residues have a root mean square deviation (r.m.s.d.) lower than a given threshold, whereas the biochemical similarity is evaluated using a Dayhoff substitution matrix (22). The score of the match is the number of matching residues which fulfil the similarity criteria. The significance of the score is evaluated by calculating the Z-score over the score distribution of the query zone comparison to the whole dataset.

THE WEB INTERFACE

The Phospho3D database can be searched by kinase name, by PDB identification code or keyword. A browsing function has been also implemented.

The information returned to the user consists of a brief description of the PDB structure(s) which fulfil the search criterion and of a list of instances presented along with
associated information (Figure 1). For each instance, the user can select three options related to the surrounding structural zone: a graphical view using the Jmol Java Applet (http://www.jmol.org); a tabular view reporting the zone annotation at the residue level; a list of 3D matches identified by local structural comparison. Each match can be visualized using Jmol. A tabular view of the matching residues is also presented (Figure 1).

CONCLUSION AND FUTURE PERSPECTIVES

The Phospho3D database is a useful tool for the analysis of the structural features of experimentally verified phosphorylation sites. Moreover, it provides the results of a large-scale local structural comparison between the zones and a representative set of eukaryotic protein chains. The results of such a comparison identify new putative phosphorylation sites and suggest the kinase(s) responsible for phosphorylation.

Phospho3D will be regularly updated as soon as the new Phospho.ELM datasets are released. The annotations will be integrated as a feature in the pdbFun server (23). We are also planning to identify and annotate those sites which are recognized by protein phosphatases and phosphoresidues-binding modules (24–26).

The Phospho3D dataset (annotations at the residue level and structural comparison results) is available upon request.

ACKNOWLEDGEMENTS

The authors would like to thank Francesca Diella for support and suggestions and for providing the Phospho.ELM dataset. We acknowledge the support of AIRC, Telethon (GGP04273), a PNR 2001–2003 (FIRB art.8) and a PNR 2003–2007 (FIRB art.8). Funding to pay the Open Access publication charges for this article was provided by AIRC.

Conflict of interest statement. None declared.

REFERENCES