Bioinformatics, 2016, 1–3
doi: 10.1093/bioinformatics/btw403
Advance Access Publication Date: 4 July 2016
Applications Note

OXFORD

Sequence analysis

MetaPred2CS: a sequence-based meta-predictor for protein–protein interactions of prokaryotic two-component system proteins

Altan Kara, Martin Vickers, Martin Swain, David E. Whitworth and Narcis Fernandez-Fuentes*

Institute of Biological, Environmental and Rural Sciences, Aberystwyth University, Aberystwyth SY23 3EB, UK

*To whom correspondence should be addressed.
Associate Editor: Alfonso Valencia

Received on February 22, 2016; revised on May 9, 2016; accepted on June 20, 2016

Abstract

Motivation: Two-component systems (TCS) are the main signalling pathways of prokaryotes, and control a wide range of biological phenomena. Their functioning depends on interactions between TCS proteins, the specificity of which is poorly understood.
Results: The MetaPred2CS web-server interfaces a sequence-based meta-predictor specifically designed to predict pairing of the histidine kinase and response-regulator proteins forming TCSs. MetaPred2CS integrates six sequence-based methods using a support vector machine classifier and has been intensively tested under different benchmarking conditions: (i) species specific gene sets; (ii) neighbouring versus orphan pairs; and (iii) k-fold cross validation on experimentally validated datasets.
Availability and Implementation: Web server at: http://metapred2cs.ibers.aber.ac.uk/, Source code: https://github.com/martinjvickers/MetaPred2CS or implemented as Virtual Machine at: http://metapred2cs.ibers.aber.ac.uk/download
Contact: naf4@aber.ac.uk
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Two-component systems (TCSs) are the most common signal transduction pathway found in prokaryotes, and they regulate crucial biological functions such as antibiotic resistance, chemotaxis, sporulation, virulence, stationary phase transition, nitrogen regulation, competence and phosphate regulation (Whitworth, 2012). TCSs function via phosphoryl group transfer between a histidine kinase (HK) and a response regulator (RR) protein. HK-RR pairing is highly specific and experimental validations are both costly and lengthy. Here, we present MetaPred2CS, a web-server that interfaces a meta-predictor designed to predict interactions between HK and RR proteins based on six sequence-based prediction methods (Kara *et al.*, 2015). Moreover, users have also access to the source code distributed on a git-hub repository or as virtual machine to execute the method on local computers, thus providing full flexibility with regards to execution parameters, reference datasets, size of alignments, etc. We also present the prediction at genome-wide scale of the TCS interactome of *Myxococcus xanthus* DK1622 as exemplar of large-scale applicability of the method.

2 Methods

2.1 Meta-predictor and performance

MetaPred2CS provides an interface with a support vector machine (Noble, 2006) meta-predictor that integrates six sequence-based prediction methods: four genome context based methods, namely gene fusion (GF) (Enright *et al.*, 1999), phylogenetic profiling (PP) (Sun *et al.*, 2005), gene neighbourhood (GN) and gene operon methods

Table 1. Benchmark results on different datasets: P+ and P- are experimentally validated interacting and non-interacting pairs of which: NP+ are neighbouring pairs and OP+ are orphan pairs. Species-specific datasets are labelled according to the given species

Dataset	Benchmarking According to Cross-validation Levels					
	5-fold		10-fold		20-fold	
	AUC	MCC	AUC	MCC	AUC	MCC
P+/P-	94.3	0.465	95.0	0.508	94.7	0.500
NP+/P-	98.8	0.639	98.4	0.639	98.8	0.634
OP+/P-	90.3	0.409	89.4	0.407	90.3	0.410
Species-Specific Dataset		Sensitivity		Specificity	Accuracy	MCC
Escherichia coli K-12 MG1655		0.82		0.86	0.85	0.607
Myxococcus xanthus DK1622		0.92		0.87	0.87	0.582
Synechocystis sp. PCC6803		0.81		0.86	0.77	0.477
Mesorhizobium loti MAFF303099		0.75		0.89	0.88	0.476

AUC: Area under the receiver-operating characteristic curve;
MCC: Matthew's correlation coefficient.

(GO) (Shoemaker and Panchenko, 2007), and two co-evolutionary methods, namely in-silico two hybrid (i2h) (Pazos and Valencia, 2002) and mirror tree (MT) methods (Pazos and Valencia, 2001). The meta-predictor has been extensively benchmarked and tested as described previously (Kara *et al.*, 2015). A summary of the prediction performance under different benchmarking scenarios is presented in Table 1.

2.2 Implementation

MetaPred2CS webs
erver is implemented using Perl, Perl_CGI and HTML, is installed on a local cluster and runs on the BioLinux operating system and uses an Apache2 web-server. This server supports all current web-browsers and does not require any additional plugins to run. Users can also opt to install MetaPred2CS locally using the fully documented source code distributed in a git-hub repository or downloading the image of a virtual machine compatible with Virtual Box (http://virtualbox.org).

2.3 Web server usage

The only requirement to utilize MetaPred2CS is the sequence of two HK and RR proteins in FASTA format (either by uploading or copy/pasting the data). Both sequences are initially compared against pre-computed predictions and scores presented if the given pair is present in the local database (Fig. 1, first submission loop). Currently, MetaPred2CS archives pre-computed predictions for the following organisms: *Escherichia coli* K-12 MG1655, *Myxococcus xanthus* DK1622, *Erwinia amylovora* ATCC 49946 and *Pseudomonas aeruginosa* UCBPP-PA14.

If the given pair is not among the pre-calculated predictions, users are directed to a second submission page (Fig. 1, Second Submission Loop) for *de novo* prediction. On this page, users can adjust parameters to their individual needs if default values are not suitable. Further information about these parameters and the usage of MetaPred2CS webserver can be found in the 'Methods' and 'Help' web pages of the server, respectively. For more flexibility users can install a local version, see above.

Upon submission, MetaPred2CS generates a unique identification code and submission web page that can be used to track progress and retrieve prediction results (bookmarking the submission web-page is recommended). The prediction progress is updated in real time and depending on prediction parameters, the running time

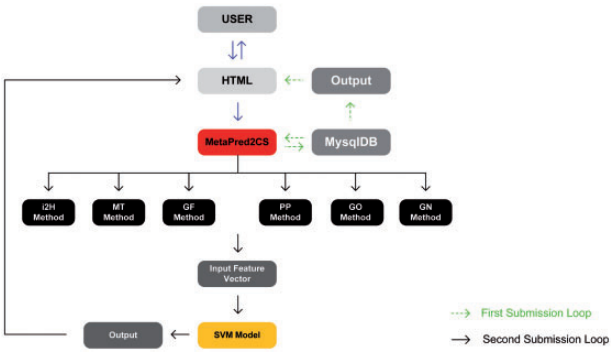

Fig. 1. Schematic representation of dataflow and components in the MetaPred2CS web-server. The paths for checking pre-calculated inputs and for de novo predictions are represented by dotted and solid arrows, respectively

is approximately 30 min (the most costly computational step is the search for protein homologs).

Upon completion, a 'Results' web page is displayed, with four different tables. The first three tables describe information on the submitted files and prediction parameters, e.g. cut-off values for each individual prediction method and a final table illustrating the prediction class: -1 and +1 for non-interacting and interacting pairs, respectively) and score, ranging from 0 to 1. Based on the score, predictions can be classified as high, medium or low confidence if the score ranges between 1.0 and 0.7; 0.7 and 0.4; and <0.4 respectively.

3 The two-component system interactome of *Myxococcus xanthus* DK1622

Several interactomes of TCS pathways (including *M.xanthus* DK1622) have been pre-computed and are available via the MetaPred2CS web-server (see before). The *M.xanthus* TCS network is poorly defined, in part because it has one of the largest complements among prokaryotes: 138 HK and 136 RR compared to 15 HK and 15 RR in *E.amylovora* and 30 HK and 32 RR in *E.coli* (Ortet *et al.*, 2015; Whitworth, 2015) (see supplementary file 1, 2, 3 for complete interactome sequences and prediction scores.) In such cases, MetaPred2CS can provide useful information on potential pairings of HKs and RRs, in turn proposing candidate biologically relevant protein complexes.

For instance, MXAN_6979-MXAN_6693(DifD) and MXAN_6692(DifE)-MXAN_6224 are two of the predicted TCS protein pairs of *M.xanthus*. DifD and DifE have a role in fibril polysaccharide production (Black and Yang, 2004). Even though the functions of MXAN_6979 and MXAN_6224 are not yet known, the predicted interactions with DifD/DifE make involvement in motility (and dependant phenomena such as fruiting body formation) plausible. MXAN_1129(FrgB)-MXAN_6099 and MXAN_4140(FrzE)-MXAN_6980 are also among the predicted protein pairs of *M.xanthus*. FrgB and FrzE regulate fruiting body formation and vegetative swarming through S motility (Cho *et al.*, 2000), suggesting that MXAN_6099 and MXAN_6980 may also regulate motility. Similarly, the predicted novel interaction pair, MXAN_0733: MXAN_3606(RodK), suggests that like RodK, MXAN_0733 may take part in the coordination of cellular aggregation during the early stages of fruiting body formation (Wegener-Feldbrügge and Søgaard-Andersen, 2009).

4 Conclusion

In this work, we present a web server for prediction of PPIs in prokaryotic TCS signalling pathways. Predictions are generated by a sequence-based meta-predictor, which is interfaced through a freely available user-friendly web-server (http://metapred2cs.ibers.aber.ac.uk/). For local use, the source code and installation instruction is available at: https://github.com/martinjvickers/MetaPred2CS; and a virtual machine implementing the method at: http://metapred2cs.ibers.aber.ac.uk/MetaPred2CS.ova. MetaPred2CS also contains precomputed predictions for a number of organisms.

Acknowledgements

We thank the P2CS team for help extracting information from the P2CS database. AK received an IBERS scholarship. NFF thanks constructive comments for anonymous reviewers.

Conflict of Interest: none declared.

References

Black,W.P. and Yang,Z. (2004) *Myxococcus xanthus chemotaxis* homologs DifD and DifG negatively regulate fibril polysaccharide production. *J. Bacteriol.*, **186**, 1001–1008.

Cho,K. *et al.* (2000) Developmental aggregation of *Myxococcus xanthus* requires frgA, an frz-related gene. *J. Bacteriol.*, **182**, 6614–6621.

Enright,A.J. *et al.* (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature*, **402**, 86–90.

Kara,A. *et al.* (2015) Genome-wide prediction of prokaryotic two-component system networks using a sequence-based meta-predictor. *BMC Bioinformatics*, **16**, 297.

Noble,W.S. (2006) What is a support vector machine? *Nat. Biotechnol.*, **24**, 1565–1567.

Ortet,P. *et al.* (2015) P2CS: updates of the prokaryotic two-component systems database. *Nucleic Acids Res.*, **43**, D536–D541.

Pazos,F. and Valencia,A. (2001) Similarity of phylogenetic trees as indicator of protein–protein interaction. *Protein Eng.*, **14**, 609–614.,

Pazos,F. and Valencia,A. (2002) In silico two-hybrid system for the selection of physically interacting protein pairs. *Proteins*, **47**, 219–227.

Shannon,P. *et al.* (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.

Shoemaker,B.A. and Panchenko,A.R. (2007) Deciphering protein–protein interactions. Part II. Computational methods to predict protein and domain interaction partners. *PLoS Comput. Biol.*, **3**, e43.

Sun,J. *et al.* (2005) Refined phylogenetic profiles method for predicting protein–protein interactions. *Bioinfo. Oxf. Engl.*, **21**, 3409–3415.

Wegener-Feldbrügge,S. and Søgaard-Andersen,L. (2009) The atypical hybrid histidine protein kinase RodK in *Myxococcus xanthus*: spatial proximity supersedes kinetic preference in phosphotransfer reactions. *J. Bacteriol.*, **191**, 1765–1776.

Whitworth,D.E. (2012) Two-component regulatory systems in prokaryotes. In: Filloux,A. (ed.) *Bacterial Regulatory Networks*. Horizon Scientific Press, Norfolk, p. 191–222.

Whitworth,D.E. (2015) Genome-wide analysis of myxobacterial two-component systems: genome relatedness and evolutionary changes. *BMC Genomics*, **16**, 780.