BIOINFORMATICS APPLICATIONS NOTE
Vol. 28 no. 3 2012, pages 426–427
doi:10.1093/bioinformatics/btr659

Sequence analysis
Advance Access publication November 29, 2011

NRPSsp: non-ribosomal peptide synthase substrate predictor

Carlos Prieto*, Carlos García-Estrada, Diego Lorenzana and Juan Francisco Martín
Institute of Biotechnology of Leon, INBIOTEC, Parque Científico de León, 24006 León, Spain
Associate Editor: Alfonso Valencia

ABSTRACT

Summary: Non-ribosomal peptide synthetases (NRPSs) are multi-modular enzymes, which biosynthesize many important peptide compounds produced by bacteria and fungi. Some studies have revealed that an individual domain within the NRPSs shows significant substrate selectivity. The discovery and characterization of non-ribosomal peptides are of great interest for the biotechnological industries. We have applied computational mining methods in order to build a database of NRPSs modules that bind to specific substrates. We have used this database to build a hidden Markov model predictor of substrates that bind to a given NRPS.

Availability: The database and the predictor are freely available on an easy-to-use website at www.nrpssp.com.

Contact: carlos.prieto@unileon.es

Supplementary information: Supplementary data is available at *Bioinformatics* online.

Received on July 12, 2011; revised on November 2, 2011; accepted on November 24, 2011

1 INTRODUCTION

Nonribosomal peptide synthetases (NRPSs) are multi-modular enzymes involved in the biosynthesis of natural products. A minimal NRPS module contains specific functional domains which are able to catalyze several activities, such as amino acid adenylation (A-activation), thioesterification (T- thiolation or acyl carrier domain) and peptide-bond formation (C-condensation domain), allowing elongation of the nascent peptide (Schwarzer and Marahiel, 2001). The primary composition of the final product is determined by the sequential order of the A-domains along the synthetase, because each A-domain recruits a particular type of substrate. The crystal structure of the peptide synthetase GrsA, which was solved with a bound, a phenylalanine substrate molecule, has enabled the identification of 10 key residues in the A-domain, which are important for the substrate binding (Conti *et al.*, 1997). Accordingly, these residues can be determinant in the substrate specificity of A-domains, and their extraction from characterized A-domains has achieved a collection of key residues signatures and general rules for deducing substrate specificity of non-characterized A-domains (Challis *et al.*, 2000; Stachelhaus *et al.*, 1999). Moreover, machine learning techniques have been applied to build a classifier based on 20 key residues and on the physico-chemical properties of amino acids to gain prediction power (Rausch *et al.*, 2005; Röttig *et al.*, 2011). Consequently, software tools and databases have been developed to collect NRPS products, such as NORINE (Caboche *et al.*, 2008) and NRPS-PKS (Anand *et al.*, 2010), and to predict the binding substrates of an NRPS such as NRPS-PKS (Ansari *et al.*, 2004), NP.Searcher (Li *et al.*, 2009), PKS/NRPS Analysis (Bachmann and Ravel, 2009) and NRPSPredictor2 (Röttig *et al.*, 2011). These prediction websites are mainly based on the generally accepted rule of analysing the active site within the A-domains. However, it is known that this approach has difficulties analysing certain types of synthetases, especially those which belong to fungi species. Problems can be caused because the GrsA crystal seems to be an inadequate model for them (Jenke-Kodama and Dittmann 2009), or because the large number of sequence variants in the active centre does not allow a correct extraction of the key residues for prediction. This observation suggests the interest of developing new prediction methods supported by other approaches. One of these approaches could be the use of hidden Markov model (HMM) as Khurana *et al.* (2010) have applied to functionally classify the acyl:CoA synthetase super-family members. This work suggests that the application of HMM profiles to classify this superfamily outperforms the predictions based on a limited number of active site residues (Khurana *et al.*, 2010). The methods stated above can also be applied to a more ambitious goal, such as the determination of the substrate that binds to an adenylation domain.

The current *omics* era has enabled the exponential growth of the sequenced NRPS. This implies that a tool which could predict the specificity of their A domains is of increasing interest, and its training could be beneficial with the new annotated NRPS. These facts, the previous experience of our group in the area, and the cited publications, have enabled the presented work, whose ultimate goal is to develop a new bioinformatic tool in order to achieve the collection, annotation, storage and prediction of substrates, which bind to adenylation domains in NRPSs. This open software tool applies a new approach in the areas of, the prediction based on HMM, enlarged training sets applying mining techniques, the regular update of its database and its design for the functional analysis of incoming NGS data.

2 METHODS

This work has been developed in three phases: (i) construction of a database with adenylation domains, which bind a known substrate; (ii) build, train and test a computational predictor; and (iii) development of a web tool. The global work flow is represented in the Supplementary Figure S1 and a detailed description of the methods is in the Supplementary Material. In order to construct the database, a semi-automatic annotation protocol was implemented and will be applied regularly in order to update the database (see Supplementary Material for a detailed description of the methods). Regarding the predictor, strategies based on position specific scoring matrices (PSSMs) and HMMs were tested to build the classifier. HMM approaches obtain better results due to the sequence heterogeneity of the A-domains and consequently the difficulty of its alignment. That is why the classification method was developed with HMM, although the idea of identifying key residues in input

*To whom correspondence should be addressed.

sequences was abandoned. A cross-validation of the predictor was done applying a leave one out test (LOO) and an receiver operating characteristic curve of these results was plotted with the R package ROCR (Supplementary Fig. S2). The input sequences of the LOO test were also analysed with NRPSpredictor2 (one class classifier) and PKS/NRPS Analysis in order to compare different approaches. The predictor is available online through the website www.nrpssp.com. It was developed with LAMP architecture (Linux, Apache, MySQL and PHP).

3 RESULTS

The large increase of sequenced proteins that has occurred in recent years has enabled the collection of more data than in previous studies. Proteins (37 126) were initially worked with, which were annotated as NRPSs or had at least one A-domain. However, only a small subset of these proteins are fully annotated (only 721 proteins are in the Swissprot subset), and a small fraction of this subset was useful for building the database. The automatic annotation of substrates with its corresponding A-domain obtained 1490 entries. Then, a data curation was manually done correcting the existing data errors and deleting the doubtful entries [mainly because (i) they do not belong to an NRPS module and (ii) the lack of knowledge of the exact correspondence with a substrate]. This process results in a database with 1598 domains, which have a known binding substrate. From these, 1578 sequences were used for training the classifier because the rest binds to substrates that have <4 annotated sequences. Although the size is not too large, it is the biggest database that has been used to train a method which predicts NRPS substrates. Rausch *et al.* (2005) used a database with 394 entries, of which 300 were used to train the SVM classifier, and the recent update of this method used a database with 557 entries (the number of entries to train the classifier was not described) (Rausch *et al.*, 2005; Röttig *et al.*, 2011). This means that our database has more than triple the size of previous ones. It is available online and the semi-automatic methods that have been developed allow its regular update. We expect that this resource will be a reference set for future research in this area.

This database was used to construct the classifier by means of the application of HMM profiles. The reliability of the classifier was measured using all data as a train and test set and by a LOO method as well (Section 2). The error rate was 13.6 and 4.96% for LOO and whole training data, respectively. If low score results (highlighted in red in the web application) are considered as not available, the error rate decreases to 5.77 and 3.76%, respectively. The LOO sequences were analysed with NRPSpredictor2 and PKS/NRPS Analysis in order to estimate their error rate in similar terms. The NRPSpredictor2 test obtained an error rate of 22.6% taking into account all the predictions and 8.29% excluding null and unavailable predictions. Similarly, PKS/NRPS Analysis obtained error rates of 73.7 and 27.35%, respectively (Supplementary Table S1). This is a very promising result, indicating that the use of more comprehensive training data and HMM achieves a more reliable predictor. In addition, the results obtained by classifying fungal proteins were studied separately. The error rate excluding low score results were 4.35% for LOO and 2.17% for whole training data, and if the low score results are not excluded, 35% for LOO and 4.10% for whole training data. An increase of the error rate is noticeable when low score results are not filtered. This increase is induced by the wide variety of fungi A-domains and their small number in the training set. Similar results have been obtained with NRPSpredictor2 and PKS/NRPS Analysis, whose predictions had a low coverage (around 30%) and an error rate of 36 and 9.1%, respectively (excluding null and non-available predictions, see Supplementary Table S2). However, coverage problems are expected to disappear as the number of fungi A-domains in the test set is increased, and this is a major objective which NRPSsp attempts to address with frequent updates.

NRPSsp is available via the website www.nrpssp.com. This website easily allows the analysis of a set of sequences, which are passed as parameters in a FASTA format (Supplementary Fig. S3 shows an example). In addition, the website has a download section which contains the updated database that has been used to train the current classifier and the HMM profiles which have been built. It enables future studies in the area and the execution of the classifier in a stand-alone mode. In this way, the application is designed for use with NGS data, which is becoming common in biotechnological research, and allows a quick functional annotation of NRPS proteins and knowledge of the substrate specificity of their A domains.

Funding: Agencia de Inversiones y Servicios de Castilla y León (record CCTT/10/LE/0001); Juan de la Cierva programme (JCI-2009-05444) of the Ministry of Science and Innovation (Spain) (to C.P.).

Conflict of Interest: none declared.

REFERENCES

Anand,S. *et al.* (2010) SBSPKS: structure based sequence analysis of polyketide synthases. *Nucleic Acids Res.*, **38,** W487–W496.

Ansari,M.Z. *et al.* (2004) NRPS-PKS: a knowledge-based resource for analysis of NRPS/PKS megasynthases. *Nucleic Acids Res.*, **32**, W405–W413.

Bachmann,B.O. and Ravel,J. (2009) Chapter 8. Methods for in silico prediction of microbial polyketide and nonribosomal peptide biosynthetic pathways from DNA sequence data. *Methods Enzymol.*, **458**, 181–217.

Caboche,S. *et al.* (2008) NORINE: a database of nonribosomal peptides. *Nucleic Acids Res.*, **36**, D326– D331.

Conti,E. *et al.* (1997) Structural basis for the activation of phenylalanine in the non-ribosomal biosynthesis of gramicidin S. *EMBO J.*, **16**, 4174–4183.

Challis,G.L. *et al.* (2000) Predictive, structure-based model of amino acid recognition by nonribosomal peptide synthetase adenylation domains. *Chem. Biol.*, **7**, 211–224.

Jenke-Kodama,H. and Dittmann,E. (2009) Bioinformatic perspectives on NRPS/PKS megasynthases: advances and challenges. *Nat. Prod. Rep.*, **26**, 874–883.

Khurana,P. et al. (2010) Genome scale prediction of substrate specificity for acyl adenylate superfamily of enzymes based on active site residue profiles. *BMC Bioinformatics*, **11**, 57.

Li,M.H. *et al.* (2009) Automated genome mining for natural products. *BMC Bioinformatics*, **10**, 185.

Rausch,C. *et al.* (2005) Specificity prediction of adenylation domains in nonribosomal peptide synthetases (NRPS) using transductive support vector machines (TSVMs). *Nucleic Acids Res.*, **33**, 5799–5808.

Röttig,M. *et al.* (2011) NRPSpredictor2–a web server for predicting NRPS adenylation domain specificity', *Nucleic Acids Res*, **39**, W362–W367.

Schwarzer,D. and Marahiel,M.A. (2001) Multimodular biocatalysts for natural product assembly. *Naturwissenschaften*, **88**, 93–101.

Stachelhaus,T. *et al.* (1999) specificity-conferring code of adenylation domains in nonribosomal peptide synthetases. *Chem. Biol.*, **6**, 493–505.