Bioinformatics, 32(3), 2016, 325–329
doi: 10.1093/bioinformatics/btv600
Advance Access Publication Date: 14 October 2015
Discovery Note

Sequence analysis

Computational discovery of specificity-conferring sites in non-ribosomal peptide synthetases

Michael Knudsen[1,2,*], Dan Søndergaard[1,2], Claus Tofting-Olesen[1,3], Frederik T. Hansen[1,3], Ditlev Egeskov Brodersen[1,3] and Christian N. S. Pedersen[1,2,*]

[1]NANORIPES – Centre for Natural Non-Ribosomal Peptide Synthesis, [2]Bioinformatics Research Centre, and [3]Department of Molecular Biology and Genetics, Aarhus University, Aarhus, Denmark

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on July 29, 2015; revised on October 11, 2015; accepted on October 12, 2015

Abstract

Motivation: By using a class of large modular enzymes known as Non-Ribosomal Peptide Synthetases (NRPS), bacteria and fungi are capable of synthesizing a large variety of secondary metabolites, many of which are bioactive and have potential, pharmaceutical applications as e.g. antibiotics. There is thus an interest in predicting the compound synthesized by an NRPS from its primary structure (amino acid sequence) alone, as this would enable an *in silico* search of whole genomes for NRPS enzymes capable of synthesizing potentially useful compounds.

Results: NRPS synthesis happens in a conveyor belt-like fashion where each individual NRPS module is responsible for incorporating a specific substrate (typically an amino acid) into the final product. Here, we present a new method for predicting substrate specificities of individual NRPS modules based on occurrences of motifs in their primary structures. We compare our classifier with existing methods and discuss possible biological explanations of how the motifs might relate to substrate specificity.

Availability and implementation: SEQL-NRPS is available as a web service implemented in Python with Flask at http://services.birc.au.dk/seql-nrps and source code available at https://bitbucket.org/dansondergaard/seql-nrps/.

Contact: micknudsen@gmail.com or cstorm@birc.au.dk

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

1 Introduction

Non-Ribosomal Peptide Synthetases (NRPS) are large proteins found mainly in bacteria and fungi where they are responsible for synthesis of a variety of peptides, many of which are bioactive and have significant pharmaceutical value (Finking and Marahiel, 2004). While the most famous example is the penicillin precursor tripeptide (synthesized by fungi in the *Penicillium* genus) many other compounds that are active as antibiotics or e.g. immunosuppressants are known. In a time where antibiotic resistance is becoming increasingly problematic, investigating the possibilities for synthesis of novel therapeutical compounds from NRPSs is therefore of great interest.

NRPSs are modular enzymes with each module comprising at least three domains: An *adenylation* (A) domain, a *thiolation* (T) domain and a *condensation* (C) domain (Finking and Marahiel, 2004). Of these, the A domain recruits a specific substrate and activates it by transfer of an adenylate moiety. The activated substrate is then picked up by a prosthetic phosphopantetheine arm coupled to the

T domain and physically moved to the C domain where the peptide bond is formed to an activated substrate from the next module in line. This process continues up until the last C domain (which in bacteria this usually is a *thioesterase* domain) where a hydrolysis reaction causes the chain of substrates to detach from the NRPS, often by circularization of the final product. One module is thus responsible for incorporating one substrate into the final peptide product.

More than 500 individual substrates have been observed in peptides synthesized by NRPSs, including both the 20 proteinogenic L-amino acids as well as their D-isomers, non-proteinogenic amino acids (such as ornithine) and fatty acids (Caboche *et al.*, 2008). Determining the substrate specificity of a given A domain biochemically by isolating it and trying all possible candidates is a cumbersome process and often fruitless effort. Hence, there is a desire to predict the specificity from the amino acid sequences of A domains alone, and state-of-the-art prediction methods such as NRPSsp (Prieto *et al.*, 2012) and NRPSpredictor2 (Röttig *et al.*, 2011) are able to achieve that with relatively high accuracy using profile Hidden Markov Models (pHMM) and support vector machines, respectively.

The first successful attempt to predict A domain substrate specificity was devised by examining the crystal structure of the A domain PheA from GrsA in *Bacillus brevis* in complex with its cognate substrate phenylalanine (PHE) (Stachelhaus *et al.*, 1999). The analysis revealed 10 residues critically involved in substrate binding and recognition. Due to high sequence similarity between A domains in general, it was speculated that sequence alignment of PheA with other A domains would reveal the residues important for substrate recognition. Using this approach, the authors successfully predicted the specificity of most A domains in their dataset using only these 10 residues, which were then aptly named the *specificity-conferring code*.

Here, we present a new approach for predicting the specificity of NRPS A domains using automatic identification of sequence motifs discriminating sequences of A domains with different specificities. Evaluated on a manually curated, non-redundant dataset, our method achieves an overall prediction accuracy of 71.3%, higher than the 66.3% obtained by the current state-of-the-art method. The motifs discovered include the well-known specificity-conferring code but also other sites not previously associated with substrate specificity. We then discuss the significance of these sites from a biological and structural perspective.

2 Methods

2.1 Data

The data used for classification consists of amino acid sequences of A domains annotated with their corresponding substrate specificities. The largest dataset available was collected for training the NRPSsp (Prieto *et al.*, 2012) classifier and contains 1578 sequences. However, as recently pointed out by Khayatt *et al.* (2013), this dataset contains many (near) duplicate sequences, incorrectly annotated A domains and even sequences not related to NRPSs. As a consequence, the authors manually curated a new dataset from public databases and merged it with data from NRPSpredictor2 (Röttig *et al.*, 2011). They then removed (near) duplicate sequences and obtained a dataset comprising 537 sequences corresponding to a total of 37 different substrate specificities. We use this dataset in this study. The dataset contains both bacterial and fungal A domain sequences, and we do not distinguish between these in training or testing of the method.

2.2 Specificity-conferring code

Using Clustal Omega (Sievers *et al.*, 2011), we generated a multiple alignment of the PheA sequence and all the sequences in the dataset, and from this extracted the positions corresponding to the specificity-conferring code in PheA as described in Stachelhaus *et al.* (1999).

2.3 Sequence Learner

We use Sequence Learner (SEQL) (Ifrim and Wiuf, 2011), an algorithm which, given two sets of sequences, identifies the sequence motifs that best discriminate between sequences from each set. In this case, we wanted to distinguish sequences of A domains binding a specific substrate S from those that do not. The training set thus consists of pairs $\{x_i, y_i\}_{i=1}^{N}$, where x_i is an A domain sequence and $y_i \in \{-1, +1\}$ indicates whether or not the corresponding A domain binds S.

The feature space (the variables on which classification is based) of SEQL is the set of all subsequences of sequences present in the training set. Let d denote the size of the feature space. Each sequence x_i can then be represented as a binary vector $x_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,d})^{\mathrm{T}}$, where $x_{i,j} \in \{0, 1\}$ indicates whether or not the jth feature is present in the sequence. Using a gradient descent algorithm, SEQL then finds parameters $\beta = (\beta_1, \beta_2, \ldots, \beta_d)$ that minimize a given loss function. Here, we use a regularized squared-hinge loss function,

$$L(\beta) = \sum_{i=1}^{N} \max(1 - y_i \beta^t x_i, 0)^2 + CR_\alpha(\beta),$$

where C is a constant, and where

$$R_\alpha(\beta) = \alpha \sum_{j=1}^{d} |\beta_j| + (1 - \alpha)\frac{1}{2}\sum_{j=1}^{d} \beta_j^2$$

is a regularizer penalizing large coordinates of β to prevent over-fitting. In our analysis, we set $C = 0.7$ and $\alpha = 0.9$. In each step, only the coordinate corresponding to the maximal gradient magnitude is updated. Even though the feature space is enormous, the search can be done efficiently by pruning the search tree: This relies on the fact that the frequency of a subsequence is bounded by the frequency of its prefixes, and hence large groups of sequences can be discarded because their prefixes are less frequent than the most frequent subsequence found so far.

It was proved in Ifrim and Wiuf (2011) that the method is guaranteed to converge. In practice, however, the algorithm is terminated when the improvement in one step is below a certain threshold. The optimization is always started from $\beta = (0, 0, \ldots, 0)^{\mathrm{T}}$, and when the process terminates, the non-zero elements in β are interpreted as weights of significant features. A positive (resp. negative) value of β_i corresponds to the ith feature being overrepresented among the positive (resp. negative) examples in the training set. When applied to A domain sequences, the SEQL method typically finds <20 significant motifs.

3 Results and discussion

To locate sequence motifs with a significant influence on A domain specificity, we used the dataset collected by Khayatt *et al.* (2013), which had been manually curated and contains no (near) duplicate sequences (see Section 2). For each substrate specificity present in the set, we built a SEQL model which discriminates between that specificity and all other specificities, and given an unknown A domain, we then apply all these models and determine the specificity to be the one corresponding to the highest scoring model. This way of constructing a multi-class classifier from a binary classifier is known as the one-*versus*-rest schema.

To evaluate the performance of SEQL, we conducted a leave-one-out (LOO) cross validation: For each sequence *S* in the dataset, we built models based on all sequences except *S*, and we then predict the specificity of *S* based on this new set of models and ask whether the highest scoring model is the one corresponding to the substrate specificity of *S*. Note that since the data contain no (near) duplicate sequences, the sequences for which we are trying to predict the specificity are never present in the corresponding training set, and hence we do not risk achieving an artificially high performance by predicting specificities already contained within the training set.

3.1 Comparison with existing methods

When evaluating SEQL on the dataset from Khayatt *et al.* (2013) using LOO an overall accuracy of 71.3% was obtained, which is higher than the 66.3% achieved by Khayatt *et al.* (2013) using an ensemble of pHMMs.

While the NRPSsp predictor (Prieto *et al.* 2012) is available as a web service, there is no option to train it on a different dataset. We are therefore unable to assess the performance of NRPSsp on the dataset used here. However, we can perform a LOO analysis using SEQL on the NRPSsp dataset, and we obtain an accuracy of 83.5%, which is comparable to the 86.4% reported by Prieto *et al.* (2012). Similarly, we are not able to train NRPSpredictor2 on the dataset from Khayatt *et al.* (2013). Note also that NRPSpredictor2 is a semi-supervised predictor: Besides training on labeled data (576 sequences of A domains with known specificities), it also takes unlabeled data into account (5096 sequences of A domains with *unknown* specificities). Since SEQL does not permit inclusion of unlabeled data, we are not able to perform a comparison of SEQL with NRPSpredictor2.

3.2 Single substrates

The accuracy of prediction for single substrates obtained using SEQL is shown in Table 1. In general, the method performs well on single substrates, but the prediction accuracy for PHE is noticeably lower than for all other substrates. This is also the case using ensemble pHMMs (see the LOO column in Table 4 in Khayatt *et al.*, 2013), and PHE also ranks among the substrates most poorly classified by NRPSpredictor2 (see Table 1 in Röttig *et al.*, 2011). Accuracies for individual substrates are not reported for NRPSsp, so we are not able to assess whether PHE also poses a problem for this classifier.

3.3 Reduced datasets

To investigate how performance depends on the dataset size, we further filtered the dataset by clustering sequences based on sequence similarity, and we performed LOO-analyses on the resulting datasets. When varying the similarity from 90 (511 sequences) to 50% (171 sequences), the accuracy decreases close to linearly from 65 to 25%. Details are provided in the Supplementary Material.

3.4 Biological significance of motifs

To investigate whether the sequence motifs discovered by SEQL have any biological significance, we identified all occurrences of motifs in sequences with specificity *S* and mapped those, via the multiple alignment, to the corresponding positions in the PheA sequence. Thus, for each position in the PheA sequence, we obtain the frequency of how many times that site is part of a motif in a sequence of specificity *S*.

Table 1. Performance of the SEQL method on individual substrates in the dataset from Khayatt *et al.* (2013)

Substrate	Count	Precision	Recall	F_1–score
Ala	46	0.48	0.65	0.56
Leu	41	0.60	0.78	0.68
Thr	34	0.91	0.94	0.93
Val	34	0.56	0.74	0.63
Ser	33	0.77	0.91	0.83
Gly	30	0.77	0.77	0.77
Cys	27	0.74	0.85	0.79
Hpg-Hpg2Cl	21	0.90	0.90	0.90
Asn	20	0.91	1.00	0.95
Pro	20	0.75	0.75	0.75
Tyr	18	0.55	0.67	0.60
Abu-Iva	17	0.79	0.65	0.71
Glu	16	0.67	0.50	0.57
Asp	15	1.00	0.73	0.85
Phe	**15**	**0.31**	**0.27**	**0.29**
Trp	14	0.64	0.50	0.56
Ile	13	0.92	0.85	0.88
Dhb-Sal	12	0.92	0.92	0.92
Orn	12	0.56	0.42	0.48
Aad	10	1.00	0.70	0.82
Dab	10	1.00	0.80	0.89
Gln	10	0.75	0.60	0.67

Only substrates occurring at least 10 times in the dataset are shown. The prediction accuracy for phenylalanine (PHE) is noticeably lower than for other substrates (highlighted in bold).

Figure 1 summarizes the results for hydrophobic aliphatic, hydrophobic aromatic and hydrophilic substrates in the dataset (this grouping of substrates was also used in the evaluation of NRPSpredictor2 in Röttig *et al.* (2011). The SEQL method clearly ranks positions around the specificity-conferring code as important, but there are also noticeable peaks around residues at positions 110, 380 and 440 (numbers correspond to the PheA reference sequence). These regions are not considered part of the specificity-conferring code, instead they correspond to three out of 10 conserved core sequences as defined in Marahiel *et al.* (1997), namely core sequences A2, A6 and A8, respectively (see Fig. 2).

The A2 core sequence is located in a central, four-stranded parallel β-sheet in the N-terminal region of the A domain and constrains torsion angles on the conserved glycine in the sequence. The motif is believed to have purely structural significance, as it is found far away from the active site and is thus not directly involved in substrate recognition. However, as this study shows, the A2 motif may still influence the substrate specificity in an indirect way (see Fig. 1).

The A6 core sequence is located close to the active site and forms part of a distorted β-bundle in the A domain, which is believed to be involved in the adenylation reaction of A domains (Pavela-Vrancie *et al.*, 1994). Photochemical labeling of 2-azido-ATP has shown that the A6 motif is involved in binding adenine or the ribose of ATP. Within the A6 motif, a highly conserved tyrosine may stack with the adenine base of ATP and the role of A6 in specificity determination may therefore relate to the size of the amino acid substrate indirectly affecting positioning the tyrosine in relation to ATP.

Finally, the A8 core sequence is located in the hinge loop connecting the N-terminal and the C-terminal subdomains of the A domain. The motif is involved in both adenylation and thioester reaction and includes a completely conserved arginine residue that interacts with the 2′ and 3′ ribose hydroxyl groups of the adenylate intermediate. Mutagenesis has shown that a single mutation of the

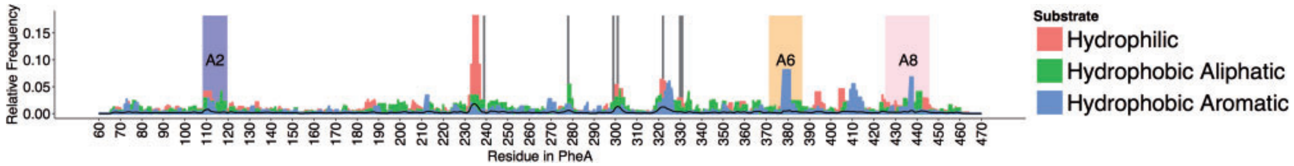

Fig. 1. Relative frequencies of how often sites in PheA correspond to sites that are part of motifs in sequences among hydrophobic aliphatic, hydrophobic aromatic and hydrophilic substrates. The black curve is the average relative frequency. Locations of the core sequences A2, A6 and A8 are shown using wide bars. The specificity-conferring sites (235, 236, 239, 278, 299, 301, 322, 330, 331) are indicated using gray bars

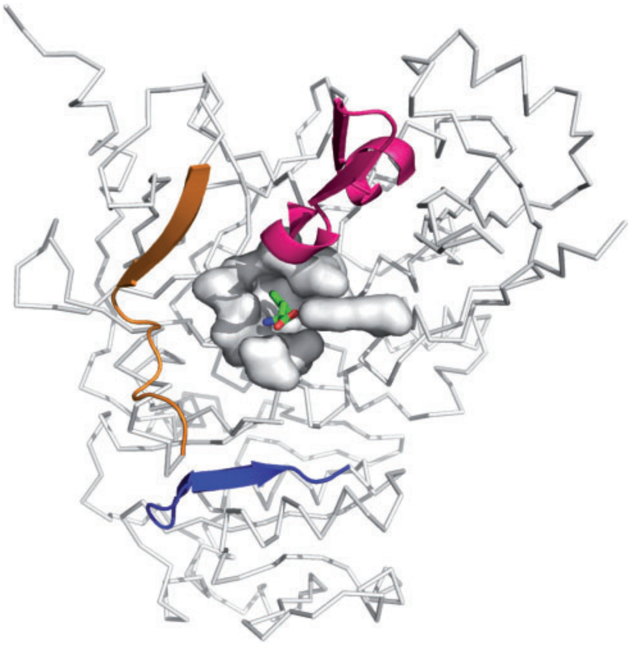

Fig. 2. Cartoon overview of PheA with the active site cleft indicated by a surface representation. Side chains of the residues constituting the specificity-conferring code and the substrate Phe are shown as stick representations. The core sequences A2, A6 and A8 are shown in blue, orange and pink, respectively

conserved glycine residue results in the inability to activate the substrate (Tokita *et al.*, 1993). The glycine residue is located in a turn between two antiparallel β-strands and this turn is of particular importance for the adenylation reaction. The A8 glycine is necessary for maintaining the turn structure between the two antiparallel β-strands. However, the exact role of the glycine residue remains unclear (Pavela-Vrancie *et al.*, 1994; Tokita *et al.*, 1993;).

The A domain adopts two major conformations during the reaction cycle: One conformation catalyzes the adenylation half-reaction while a 140° rotation orients the C-terminal subdomain of the A domain in the correct position for the second half-reaction, thioester formation (Drake *et al.*, 2006). Thus, alternate faces of the C-terminal domain are exposed to the same active site for the two half-reactions, a phenomenon referred to as domain alternation (Bandarian *et al.*, 2002). The rotation of the C-terminal subdomain moves the A8 loop into the active site responsible for the adenylation half-reaction move the conserved glycine of A8 $\sim$ 30Å. In this new position, A8 can interact with the adenylate intermediate formed by displacement of pyrophosphate. Furthermore, mutagenesis has shown that residue composition of the A8 core sequence is important for the thioester reaction (Reger *et al.*, 2007). Mutating the conserved glycine to leucine disrupts only the second half-reaction, which leads to the conclusion that the bulky side chain of leucine causes steric interference (Drake *et al.*, 2006; Reger *et al.*, 2007). The role of A8 in determining substrate specificity is therefore most likely related to positioning of the ATP and phosphopantetheine in the adenylation reaction and thioester reactions. Since neither A2, A6 nor A8 interacts directly with the amino acid substrate, we propose that the overall residue composition of these conserved core sequences influence substrate specificity indirectly.

4 Web service

Our classifier is available as a web service where users can paste or upload A domains in FASTA format (http://services.birc.au.dk/seql-nrps). The sequences are then classified and the results page shows the sequence identifier, the predicted substrate specificity, and the probability given by SEQL that the sequence belongs to the predicted substrate. This probability is colored in accordance with the confidence of the prediction, which is computed as the number of standard deviations that the highest probability is over the mean of the probabilities given by each substrate model.

The sequence is shown with motifs which discriminate for the predicted substrate colored green, and motifs discriminating against the predicted substrate colored red.

The web service, code for running the experiments and instructions for usage are available online (MIT licensed) at https://bitbucket.org/dansondergaard/seql-nrps/.

5 Conclusion

We have presented a new tool for predicting substrate specificities of NRPS adenylation domains. Besides achieving a high accuracy, the predictor differs from existing methods in that it identifies sequence motifs, which are not necessarily directly related to the active sites of the binding pocket, yet still distinguish between different substrate specificities. These motifs may be used as starting points for further investigations of the biological nature of specificity determination. Indeed, we identify three regions of which two have previously been suggested to play a role in substrate specificity determination. The predictor is available online as a free-to-use web service (http://services.birc.au.dk/seql-nrps).

Funding

Aarhus University Research Foundation (the NANORIPES project) and Danish National Research Foundation (the PUMPkin center of excellence).

Conflict of Interest: none declared.

References

Bandarian,V. *et al.* (2002) Domain alternation switches b12-dependent methionine synthase to the activation conformation. *Nat. Struct. Mol. Biol.*, **9**, 53–56.

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

Drake,E.J. *et al*. (2006) Structure of the entb multidomain nonribosomal peptide synthetase and functional analysis of its interaction with the ente adenylation domain. *Chem. Biol.*, **13**, 409–419.

Finking,R., and Marahiel,M.A. (2004) Biosynthesis of nonribosomal peptides. *Annu. Rev. Microbiol.*, **58**, 453–488.

Ifrim,G., and Wiuf,C. (2011) Bounded coordinate-descent for biological sequence classification in high dimensional predictor space. In: *Proceedings of the 17th ACM SIGKDD Conference 2011*.

Khayatt,B.I. *et al*. (2013) Classification of the adenylation and acyl-transferase activity of NRPS and PKS systems using ensembles of substrate specific hidden Markov models. *PLoS One*, **8**, e62136.

Marahiel,M.A. *et al*. (1997) Modular peptide synthetases involved in nonribosomal peptide synthesis. *Chem. Rev.*, **97**, 2651–2674.

Pavela-Vrancie,M. *et al*. (1994) Identification of the ATP binding site in tyrocidine synthetase 1 by selective modification with fluorescein 5′-isothiocyanate. *J. Biol. Chem.*, **269**, 14962–14966.

Prieto,C. *et al*. (2012) NRPSsp: non-ribosomal peptide synthase substrate predictor. *Bioinformatics*, **28**, 426–427.

Reger,A.S. *et al*. (2007) Biochemical and crystallographic analysis of substrate binding and conformational changes in acetyl-CoA synthetase. *Biochemistry*, **46**, 6536–6546.

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

Sievers,F. *et al*. (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega.*Mol. Syst. Biol.*, **7**, 539.

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

Tokita,K. *et al*. (1993) Effect of single base substitutions at glycine-870 codon of gramicidin s synthetase 2 gene on proline activation. *J. Biochem.*, **114**, 522–527.