J Ind Microbiol Biotechnol (2014) 41:461–467
DOI 10.1007/s10295-013-1322-2

ORIGINAL ARTICLE

SIMB
Society for Industrial Microbiology and Biotechnology

Predicting substrate specificity of adenylation domains of nonribosomal peptide synthetases and other protein properties by latent semantic indexing

Damir Baranašić · Jurica Zucko · Janko Diminic · Ranko Gacesa · Paul F. Long · John Cullum · Daslav Hranueli · Antonio Starcevic

Received: 17 May 2013 / Accepted: 3 August 2013 / Published online: 9 October 2013

Abstract Successful genome mining is dependent on accurate prediction of protein function from sequence. This often involves dividing protein families into functional subtypes (e.g., with different substrates). In many cases, there are only a small number of known functional subtypes, but in the case of the adenylation domains of nonribosomal peptide synthetases (NRPS), there are >500 known substrates. Latent semantic indexing (LSI) was originally developed for text processing but has also been used to assign proteins to families. Proteins are treated as ''documents'' and it is necessary to encode properties of the amino acid sequence as ''terms'' in order to construct a term-document matrix, which counts the terms in each document. This matrix is then processed to produce a document-concept matrix, where each protein is represented as a row vector. A standard measure of the closeness of vectors to each other (cosines of the angle between them) provides a measure of protein similarity. Previous work encoded proteins as oligopeptide terms, i.e. counted oligopeptides, but used no information regarding location of oligopeptides in the proteins. A novel tokenization method was developed to analyze information from multiple alignments. LSI successfully distinguished between two functional subtypes in five well-characterized families. Visualization of different ''concept'' dimensions allows exploration of the structure of protein families. LSI was also used to predict the amino acid substrate of adenylation domains of NRPS. Better results were obtained when selected residues from multiple alignments were used rather than the total sequence of the adenylation domains. Using ten residues from the substrate binding pocket performed better than using 34 residues within 8 Å of the active site. Prediction efficiency was somewhat better than that of the best published method using a support vector machine.

Keywords LSI · NRPS · Adenylation domains · Protein tokenization · Functional subtype

D. Baranašić · J. Zucko · J. Diminic · R. Gacesa · D. Hranueli · A. Starcevic (✉)
Faculty of Food Technology and Biotechnology, University of Zagreb, Pierottijeva 6, 10000 Zagreb, Croatia
e-mail: astar@pbf.hr

D. Baranašić · J. Cullum
Department of Genetics, University of Kaiserslautern, Postfach 3049, 67653 Kaiserslautern, Germany

P. F. Long
Institute of Pharmaceutical Science, King's College London, Franklin–Wilkins Building, 150 Stamford Street, London, SE1 9NH, UK

P. F. Long
Department of Chemistry, King's College London, Franklin–Wilkins Building, 150 Stamford Street, London SE1 9NH, UK

Introduction

The rapid progress in DNA sequencing technology is resulting in many new genome sequences. However, characterization of proteins or secondary metabolites from a particular strain still requires considerable effort. Genome mining aims to bridge the gap between DNA sequences and laboratory experiments by using bioinformatics to focus experiments on a limited number of interesting targets rather than using a random screening approach. Effective genome mining requires accurate identification of protein function from the sequence. In most cases, hidden Markov model profiles [6] can assign proteins to a family,

but the exact function of the encoded protein remains unknown; many families have subfamilies with different substrates or other properties. Often, there is only a small number of different known subfamilies, and methods have been developed to identify amino acid residues and predict the subfamily [7, 8].

Predicting chemical structures of secondary metabolites from DNA sequences of their synthesis genes is a difficult problem, as the synthesis pathways often involve many reactions, all of which need to be accurately predicted. Most progress has been made for modular biosynthetic clusters: polyketide synthases (PKS), nonribosomal peptide synthetases (NRPS), and mixed clusters containing both PKS and NRPS modules. These modules synthesize their products in successive steps, and each step is carried out by a different module of the enzyme. It is usually assumed that the modules function independently of each other so that the chemical structure of the product can be predicted by predicting the function of each module.

In PKS clusters, substrate specificity is determined by the acyl transferase (AT) domain of each extender module. There are five known substrates, which can be predicted well by using amino acid fingerprints of specificity-determining residues [14]. Despite the limited number of substrates, PKS modules exhibit a much larger repertoire of extension reactions because of the presence of reduction domains. There are four possible degrees of reduction and different stereochemistries encoded by the reduction domains. The program ClustScan was developed to integrate the analysis of AT-domain specificity with reduction domain functions and to predict chemical structures of modular PKS products [15].

In contrast to modular PKS clusters, most diversity in NRPS clusters is encoded by substrate choice. The adenylation domains (A) choose the specific amino acid to incorporate at each elongation reaction, but the problem of predicting substrate is difficult, because there is a large number of known substrates: ~500 [16]. An important step in predicting substrates of A domains was the identification of eight or ten critical binding-pocket residues, which correlate well with substrate specificity [2, 14]. However, there are problems in using such information for substrate prediction when there are a large number of potential substrates. In particular, it is difficult to handle sequences, which do not have a precise match in the training data. This situation prompted an alternative approach based on support-vector machines (SVM). A set of 34 amino acid residues was identified near the active center of A domains and a coding scheme developed based on their physicochemical properties [12]. The initial approach could only assign amino acids to classes of amino acids, but refinement of the approach also allowed prediction of specific amino acids [13]. As SVMs are designed to give a binary split of data sets, it was necessary to use several splits to achieve good results.

A term-document matrix was constructed using tokens as "terms" and proteins as ''documents''. Latent semantic indexing (LSI) was initially developed for analyzing sets of documents [4]. A term-document matrix is constructed that counts how often each word is present in each document of a collection. A standard theorem of linear algebra shows that the matrix can be decomposed into a product of three matrices (singular-value decomposition). The middle matrix is a diagonal matrix with the singular values in decreasing order, which can be interpreted as ''concepts,'' with the size of the value corresponding to the ''significance'' of the concept. In LSI, only the largest of the singular values are retained (typically, ~100). This is designed to reduce background noise and computational burden. The third matrix, the document-concept matrix, has rows corresponding to the documents, i.e. document vectors. Similarity of the vectors (often measured using the standard measure of the cosine of the angle between them) indicates that the documents are related. If a suitable coding is used, proteins can be viewed as documents defined by their amino acid sequences. LSI has been used to assign proteins to families [3]. These investigations used counts of single amino acids, dipeptides, and tripeptides as "terms". Generally, the best results were with tripeptides, where 8,000 combinations are possible. This approach uses counts of peptides but not their position in the protein. Although this approach loses information, it may have advantages for families of distant proteins, where multiple alignment is difficult.

This paper examines the use of LSI for assigning proteins to subfamilies. We wanted to incorporate information from multiple alignments, so we developed a new method of constructing terms based on tokenization of residues in multiple alignments. Initially, the method was applied to several well-characterized families with two subfamilies. Subsequently, the method was used to predict substrate specificities of A domains.

Materials and methods

Protein sequences

Nucleotidyl cyclase, protein kinase, lactate/malate (LDH/MDH) dehydrogenase, and ketoreductase (KR) sequences were as in Goldstein et al. [7]. The AT domain sequences were extracted from the ClustScan database [5]. The 8 Å sequences (34 amino acid residues), binding-pocket sequences (ten amino acid residues), and truncated sequences containing active site residues (136–150 amino acid residues) from 397 NRPS A domains were downloaded from

Rausch et al. [12] (supplementary material). Accession numbers for protein sequences containing A domains were downloaded from the same source and used to retrieve protein sequences from the National Center for Biotechnical Information (NCBI) GenBank database [11]. Complete A-domain sequences were defined using HMMER version 3.0 [6] with a specially generated profile.

LSI

LSI was carried out using MATLAB® [10] with the Bioinformatics Toolbox. Term-document matrices were constructed using protein sequences as documents. In some cases, mono-, di- and tri-peptides were used as "terms," as in Couto et al. [3]. In the case of using tokens as terms, multiple alignments of the proteins were constructed with Clustal X2 version 2.1 [9] using default settings. The tokens were generated by combining the amino acid single-letter code with the column number in the multiple alignment (e.g., an alanine residue at position 17 would be tokenized as 'A17'). In some cases, the log entropy global weighting function was used to modify the term-document matrix. The MATLAB function svd was used to perform a full singular value decomposition of the term-document matrix. Singular values were selected with which to choose the number of dimensions to use for LSI, with a relative variance criterion: the singular value S_i is included if:

$$\frac{S_{\mathrm{i}}^2}{\sum S_{\mathrm{i}}^2} > \frac{0.7}{n}$$

where n is the number of protein sequences. The svds function was used to calculate a decomposition with reduced dimension, i.e., LSI. Two- and three-dimensional projections of the document-concept matrix were viewed using MATLAB.

A test protein was processed in the same way as proteins in the term-document matrix. The resulting vector was processed using the folding-in method to produce a pseudovector [4]. Cosines of the angle between the pseudovector and each document vector in the document-concept matrix were calculated, and the functional subtype of the protein with the highest score was used to predict the subtype of the test protein.

Statistics for A-domain prediction

Statistical measures were as in Röttig et al. [13]. For each amino acid substrate, the number of true positives (TP), false positives (FP), and false negatives (FN) were calculated. Precision was calculated as TP/(TP + FP); recall was calculated as TP/(TP + FN). The F measure is the harmonic mean of precision and recall.

BLAST analysis

The Basic Local Alignment Search Tool Plus (BLAST+) package [1] was used. BLAST databases of complete A domains, truncated A domains, and extracted binding-pocket residues were constructed and queried with the complete set of corresponding sequences. BLAST reports were processed in MATLAB to generate the best hit, which was used to predict the amino acid substrate of A domains.

Results

Protein families with two functional subtypes

Applicability of LSI to predict functional subtypes was tested on five different protein families, each with two functional subtypes. We used a novel tokenization method to generate terms, which starts with a multiple sequence alignment. Tokens are constructed in the following way: the code corresponding to the amino acid was combined with the column number in the multiple alignment; e.g., if the amino acid residue in column 17 of the multiple sequence alignment was an alanine, the token would be 'A17'. The term-document matrix was constructed by using tokens as terms and proteins as documents. A standard log-entropy global weighting scheme was used to modify the term-document matrix. It was then used for LSI, with dimension reduction determined by a standard relative variance criterion (this resulted in 28–81 dimensions, depending on the protein family). Prediction accuracy was tested using a ''leave-one-out'' test, i.e., a series of LSI matrices were constructed, with each lacking one protein. A pseudovector corresponding to the omitted protein was then generated and compared to vectors for all other proteins by calculating the cosine of the angle between them. The closest vector (i.e., the highest cosine value) was used to predict the substrate. Table 1 shows that LSI provides good subfamily prediction: as with several other methods [7, 8], it gives almost perfect prediction of nucleotidyl cyclase substrates, distinguishes between tyrosine and serine/threonine kinases, and allows good prediction of lactate and malate dehydrogenases. It also provides a comparable performance on the other two families.

Proteins correspond to documents in the LSI analysis and are present in a document-concept matrix. LSI ranks the concepts in order of singular values associated with them, i.e., signal strength in the data. It is possible to explore whether concepts correspond to interesting differences between members of the families by viewing specific dimensions (i.e., components of proteins in the document-concept matrix). Fig. 1 shows such visualization for the

Table 1 Summary of results for specificity prediction of five different protein families with two substrate specificities [7, 8]

Protein family	This paper	Goldstein et al. [7].	Hannenhalli et al. [8]
Nucleotidyl cyclases	4/75	0/75	0/72
Protein kinases	2/215	0/215	0/293
Malate/lactate dehydrogenases	6/183	4-6/183	0/103
Ketoreductase domains	12/72	9-20/72	
Acyl transferase domains	25/610	2-5/181	

In each case, the number of false predictions (numerator) and the total number of proteins used (denominator) are given. Results are compared with those of two published papers

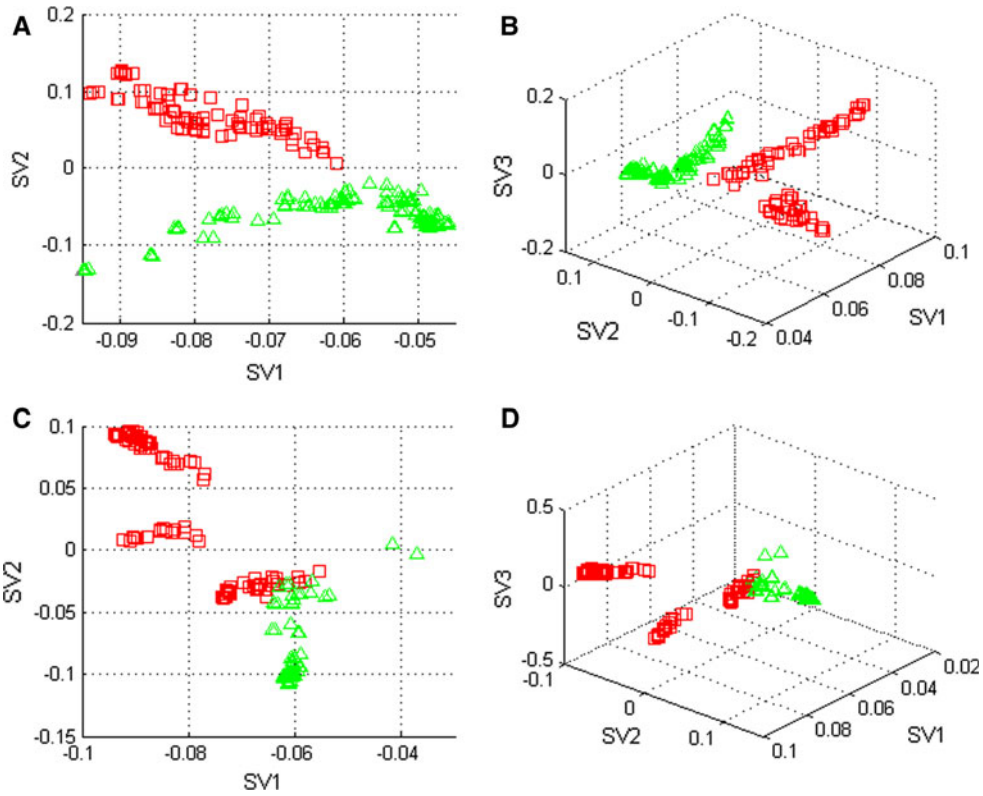

Fig. 1 Protein families using projections in the two or three dimensions corresponding to the largest singular values. **a**, **b** Protein kinases: serine–threonine kinases (*red*) and tyrosine kinases (*green*). **c**, **d** Malate and lactate dehydrogenases (*red* and *green*, respectively)

kinase and dehydrogenase families in two- and three dimensions, taking the first three dimensions corresponding to the highest singular values. It can be seen that much of the kinase family separation is achieved in the second dimension (value on the *Y* axis in Fig. 1a). In three dimensions, the serine–threonine kinases seem to split into two clusters. In the case of dehydrogenases, the MDH subfamily is split into three clusters, one of which is not well separated from the LDH subfamily in the first three dimensions (Fig. 1 c, d). However, Table 1 shows that separation is achieved when all 56 dimensions are used.

Prediction of A-domain substrates

A collection of 397 A-domain sequences [12] representing 47 different substrate specificities was used to test the applicability of LSI for predicting substrate specificity. Prediction precision was evaluated using a 'leave-one-out' test. This was done for all 383 domains, the specificity of which was present at least twice in the collection. These consisted mostly of bacterial sequences (348/383), with most of the rest being fungal sequences. Initially, all domain sequences were used. Apart from the tokenization method described above, we also used mono-, di- and tri-peptides, as in Couto et al. [3]. The percentage of domains with a correct prediction was calculated. The tokenization method was slightly better than the tripeptide method (Table 2), and dipeptides and monopeptides were considerably worse. We also compared these results by constructing a BLAST database of A-domain sequences, with the best BLAST hit used as a predictor. This showed that BLAST was marginally better than LSI methods (Table 2).

Table 2 Prediction of A-domain substrates from whole and truncated sequences

Method	Accuracy (%)	
	Whole A domain	Truncated A domain
LSI tokenization	84	87
LSI tripeptide	82	87
LSI dipeptide	76	83
LSI monopeptide	49	69
BLAST	85	87

LSI latent semantic indexing, *BLAST* Basic Local Alignment Search Tool

A-domain sequences are about 400 amino acids long on average. Stachelhaus et al. [14] found ten residues around the binding pocket, which are important for determining substrate specificity. To reduce noise, we also used truncated A-domain sequences, which spanned these ten residues; these truncated sequences are 136–150 residues long. When only these residues are used, all methods gave improved prediction (Table 2).

When we used LSI with the tokenization method restricted to the ten binding-pocket residues, there was an 89 % prediction accuracy, whereas the tripeptide coding only gave an 85 % accuracy. Challis et al. [2] used BLAST with eight critical binding-pocket residues to predict substrate specificity. The ten binding-pocket residues were used for such a BLAST analysis and gave a 79 % prediction accuracy. We also used 34 residues, which are within 8 Å of the active site [12] for LSI analysis using tokenization. However, these gave lower prediction accuracy (87 %) than the ten binding-pocket residues. Most A-domain sequences are bacterial (348/383). Bacterial and fungal specificities were predicted with 89 % and 85 % accuracy, respectively. As the number of fungal sequences is low and any differences in accuracy are not large, bacterial and fungal sequences were treated together for detailed analysis.

A detailed analysis of the prediction for each amino acid substrate was carried out. Table 3 shows the results for each amino acid substrate used by Röttig et al. [13]; this corresponds to 364 of the 397 A domains. Recall measures the proportion of A domains for each amino acid substrate correctly predicted; i.e., recall will be low if the method does not recognize the correct substrate. Precision measures the proportion of A domains predicted to use a particular substrate and that actually use the substrate; i.e., precision will be low if there are many false predictions of the substrate. In most cases, recall and precision are similar in value. In the case of isovaline (IVA), a rare amino acid incorporated by some fungal A domains, recall is 1.0 whereas precision is only 0.5; the seven cases of A domains with IVA substrate were all correctly predicted, but there were also seven other domains incorrectly predicted as IVA (two ALA, one GLY, one LEU, one SER, two VAL). In comparison, the SVM method [13] with the same seven IVA A domains had lower recall (0.73) but better precision (0.93). The F measure is the harmonic mean of recall and precision and is a convenient single measure of prediction quality. Table 3 shows a comparison of LSI results with the SVM method of Röttig et al. [13]. The average value of the F measure is 0.87, which is better than the value of 0.81 reported by Röttig et al. [13].

The A-domain prediction was implemented as a web-based program (URL bioserv7.bioinfo.pbf.hr/LSIpredictor). The sequence can be pasted into a window, and the program shows the distance of the A domain being queried to those of known substrates. This not only gives a prediction but allows the user to assess the quality of the prediction.

Discussion

LSI has been used successfully to assign proteins to protein families [3]. In this study, we explored the possibility of using LSI to distinguish functional subtypes of a family. We wanted to include information from multiple alignments, so we devised a simple tokenization scheme. Initially, five well-characterized families with two subtypes were studied. LSI successfully predicted a subtype, with similar performance to other methods. In all these cases, LSI was used with standard parameters; it is usually possible to obtain small gains in performance by optimizing the weighting scheme and the number of dimensions used. With LSI, it is possible to visualize separation of members of a protein family in two or three dimensions and to choose different dimensions for viewing (Fig. 1). This offers scope for exploring family clustering with the aim of discovering new properties. The dimensions correspond to concepts in text processing, but their significance in a protein sequence concept is almost unexplored. In the context of genome mining, it is possible that members of a family, which are not closely clustered with well-studied proteins, will prove to have interesting novel properties.

NRPS A domains offer an interesting system for testing LSI, because there are large numbers of substrates (~500; [16]). We used 397 well-characterized A domains [12]. Whole sequences of the domains are about 400 amino acids long, and we also used shorter truncated sequences (136–150 amino acids long) containing binding-pocket residues [14]. Tokenization coding performed slightly better than the tripeptide coding shown in Couto et al. [3], and truncated sequences gave improved prediction compared with whole sequences, probably

Table 3 Prediction efficiency of substrate specificities of A domains

Substrate	Recall	Precision	F measure	F measure, Röttig et al. [13]
2-amino-adipic acid (AAD)	1.00	1.00	1.00	1.00
Alanine (ALA)	0.82	0.90	0.86	0.88
Arginine (ARG)	0.80	1.00	0.89	0.83
Asparagine (ASN)	0.93	1.00	0.96	0.94
Aspartic acid (ASP)	0.83	0.91	0.87	0.70
β-hydroxy-tyrosine (BHT)	1.00	1.00	1.00	0.72
Cysteine (CYS)	1.00	0.96	0.98	1.00
2,3-dihydroxy-benzoic acid (DHB)	0.93	0.88	0.90	0.95
3,5-dihydroxy-phenyl-glycine (DHPG)	1.00	1.00	1.00	0.94
Glutamine (GLN)	1.00	0.80	0.89	0.69
Glutamic acid (GLU)	0.92	0.92	0.92	0.70
Glycine (GLY)	0.83	1.00	0.91	0.91
4-hydoxy-phenyl-glycine (HPG)	0.95	0.90	0.92	0.97
Isoleucine (ILE)	1.00	0.92	0.96	0.92
Isovaline (IVA)	1.00	0.50	0.67	0.81
Leucine (LEU)	0.81	1.00	0.89	0.78
Lysine (LYS)	0.80	1.00	0.89	0.40
Ornithine (ORN)	1.00	0.83	0.91	0.93
Phenylalanine (PHE)	0.45	0.50	0.48	0.69
Pipecolic acid (PIP)	0.60	1.00	0.75	0.70
Proline (PRO)	0.75	0.86	0.80	0.76
Serine (SER)	0.95	0.95	0.95	0.96
Threonine (THR)	0.83	0.95	0.89	0.95
Tryptophan (TRP)	0.67	1.00	0.80	0.32
Tyrosine (TYR)	0.71	0.63	0.67	0.70
Valine (VAL)	0.93	0.89	0.91	0.80

For each amino acid substrate, the number of true positives (TP), false positives (FP) and false negatives (FN) were calculated: recall = TP/(TP + FN), precision = TP/(TP + FP). The F measure is the harmonic mean of precision and recall. F-measure values from Röttig et al. [13] are given as a comparison

because there was less noise. However, LSI did not perform better than a simple BLAST analysis. We therefore decided to use residues, which are closely associated with the active site. The binding pocket was defined by ten residues, which were important for substrate selection [14]. There were also 34 residues, which were chosen as lying within 8Å of the active site [12]. LSI performed better with the ten binding-pocket residues than with the 34, 8-Å-distance residues. This is in contrast to results with SVMs, where the 8-Å residues performed better [12, 13]. Most A domains (348/383) were bacterial. However, specificity prediction of fungal domains was only slightly worse than that of bacterial domains (85 % vs 89 %, respectively). It is likely that using a larger training data set of A domains with more fungal sequences would lead to improved prediction results. A detailed comparison of prediction quality with those of Röttig et al. [13] showed that LSI performed slightly better than SVMs (Table 3). Thus, LSI is the most accurate method currently available for predicting substrate specificities of A domains. As NRPS clusters often contain many modules, even small improvements in predicting single modules can make a significant difference to predicting the final product.

A major advantage of the LSI approach is that it is easy to add further substrates to the prediction procedure. It is only necessary to align sequences and recalculate matrices. With SVMs, it is necessary to redefine sequences of binary splits to be used when new substrates are added. In the LSI approach, the cosine of the angle between vectors gives a measure of how close the predicted domain is to a domain of known substrate. This gives information about the quality of prediction, as low values indicate no close relatives and an increased risk of an FP prediction. This information is likely to be important in genome mining. In some cases, it might be important for particular amino acids to be present in the compounds of interest so that information about quality of prediction would allow some clusters to be rejected. Alternatively, the presence of an A domain with a low quality of prediction might indicate the presence of a rare amino acid and the possibility of a compound with novel properties.

The tokenization scheme used here is simple but could be supplemented by further tokens. For instance, it would

be possible to add tokens in which groups of amino acids with similar properties are given the same label (e.g., for charge or bulk).

Acknowledgments This work was supported by the grant 09/5 (to DH) from the Croatian Science Foundation, Republic of Croatia, and by a cooperation grant of the German Academic Exchange Service (DAAD) and the Ministry of Science, Education and Sports, Republic of Croatia (to DH and JC). It was also supported by King's College London (to PFL).

References

1. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL et al (2009) BLAST + : architecture and applications. BMC Bioinf 10:421
2. Challis GL, Ravel J, Townsend CA (2000) Predictive, structure-based model of amino acid recognition by nonribosomal peptide synthetase adenylation domains. Chem Biol 7:211–224
3. Couto BR, Ladeira AP, Santos MA (2007) Application of latent semantic indexing to evaluate the similarity of sets of sequences without multiple alignments character-by-character. Genet Mol Res 6:983–999
4. Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R et al (1990) Indexing by latent semantic analysis. J Am Soc Inform Sci 41:391–407
5. Diminic J, Zucko J, Trninic Ruzic I, Gacesa R, Hranueli D, Long PF, Cullum J, Starcevic A et al (2013) Databases of the Thio-template Modular Systems (*CSDB*) and their in silico recombinants (*r-CSDB*). J Ind Microbiol Biotechnol 40:653–659
6. Eddy SR (2008) A probabilistic model of local sequence alignment that simplifies statistical significance estimation. PLoS Comput Biol 4:e1000069
7. Goldstein P, Zucko J, Vujaklija D, Krisko A, Hranueli D, Long PF, Etchebest C, Basrak B, Cullum J et al (2009) Clustering of protein domains for functional and evolutionary studies. BMC Bioinformatics 10:335
8. Hannenhalli SS, Russell RB (2000) Analysis and prediction of functional sub-types from protein sequence alignments. J Mol Biol 303:61–76
9. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG et al (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23:2947–2948
10. MATLAB®. http://www.mathworks.com/products/matlab/
11. National Center for Biotechnology Information (NCBI GenBank database). http://www.ncbi.nlm.nih.gov/
12. Rausch C, Weber T, Kohlbacher O, Wohlleben W, Huson DH 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
13. Röttig M, Medema MH, Blin K, Weber T, Rausch C, Kohlbacher O (2011) NRPSpredictor2–a web server for predicting NRPS adenylation domain specificity. Nucleic Acids Res (Web Server issue) 39:W362–W367
14. Stachelhaus T, Mootz HD, Marahiel MA (1999) The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases. Chem Biol 6:493–505
15. Starcevic A, Zucko J, Simunkovic J, Long PF, Cullum J, Hranueli D (2008) ClustScan: an integrated program package for the semi-automatic annotation of modular biosynthetic gene clusters and in silico prediction of novel chemical structures. Nucleic Acids Res 36:6882–6892
16. Strieker M, Tanovic A, Marahiel MA (2010) Nonribosomal peptide synthetases: structures and dynamics. Curr Opin Struct Biol 20:234–240