Nucleic Acids Research, 2004, Vol. 32, Web Server issue ***W405–W413***
DOI: 10.1093/nar/gkh359

NRPS-PKS: a knowledge-based resource for analysis of NRPS/PKS megasynthases

Mohd. Zeeshan Ansari, Gitanjali Yadav, Rajesh S. Gokhale and Debasisa Mohanty*

National Institute of Immunology, Aruna Asaf Ali Marg, New Delhi 110067, India

Received January 18, 2004; Revised and Accepted February 19, 2004

ABSTRACT

NRPS-PKS is web-based software for analysing large multi-enzymatic, multi-domain megasynthases that are involved in the biosynthesis of pharmaceutically important natural products such as cyclosporin, rifamycin and erythromycin. NRPS-PKS has been developed based on a comprehensive analysis of the sequence and structural features of several experimentally characterized biosynthetic gene clusters. The results of these analyses have been organized as four integrated searchable databases for elucidating domain organization and substrate specificity of nonribosomal peptide synthetases and three types of polyketide synthases. These databases work as the backend of NRPS-PKS and provide the knowledge base for predicting domain organization and substrate specificity of uncharacterized NRPS/PKS clusters. Benchmarking on a large set of biosynthetic gene clusters has demonstrated that, apart from correct identification of NRPS and PKS domains, NRPS-PKS can also predict specificities of adenylation and acyltransferase domains with reasonably high accuracy. These features of NRPS-PKS make it a valuable resource for identification of natural products biosynthesized by NRPS/PKS gene clusters found in newly sequenced genomes. The training and test sets of gene clusters included in NRPS-PKS correlate information on 307 open reading frames, 2223 functional protein domains, 68 starter/extender precursors and their specific recognition motifs, and also the chemical structure of 101 natural products from four different families. NRPS-PKS is a unique resource which provides a user-friendly interface for correlating chemical structures of natural products with the domains and modules in the corresponding nonribosomal peptide synthetases or polyketide synthases. It also provides guidelines for domain/module swapping as well as site-directed mutagenesis experiments to engineer biosynthesis of novel natural products. NRPS-PKS can be accessed at http://www.nii.res.in/nrps-pks.html.

INTRODUCTION

Nonribosomal peptide synthetases (NRPSs) and polyketide synthases (PKSs) are multi-enzymatic, multi-domain megasynthases involved in the biosynthesis of nonribosomal peptides and polyketides. These secondary metabolites exhibit a remarkable array of biological activity and many of them are clinically valuable anti-microbial, anti-fungal, anti-parasitic, anti-tumor and immunosuppressive agents (1–3). Nonribosomal peptides are biosynthesized by sequential condensation of amino acid monomers, whereas polyketides are made from repetitive addition of two carbon ketide units derived from thioesters of acetate or other short carboxylic acids. NRPSs and modular PKSs are comprised of so-called modules, which are sets of distinct active sites for catalysing each condensation and chain elongation step (4–9). Each module in an NRPS or PKS consists of certain obligatory or core domains (Supplementary Figure 1) for addition of each peptide or ketide unit and a variable number of optional domains responsible for modification of the peptide/ketide backbone. The minimal core module in the case of an NRPS consists of an adenylation (A) domain for selection and activation of amino acid monomers, a condensation (C) domain for catalysing the formation of peptide bonds and a thiolation or peptidyl carrier protein (T or PCP) domain with a swinging phosphopantetheine group for transferring the monomers/growing chain to various catalytic sites. Similarly, an acyltransferase (AT) domain for extender unit selection and transfer, an acyl carrier protein (ACP) with a phosphopantetheine swinging arm for extender unit loading and a ketoacyl synthase (KS) domain for decarboxylative condensations constitute the core domains of PKS modules (5–9). During the biosynthesis, the growing chain remains covalently attached to the enzyme and upon reaching its full length, a thioesterase (TE) domain catalyses the release of the NRPS and PKS products. The segments of polypeptide chain connecting all these domains are referred to as linkers and they have been shown to establish functional communication between and within modules (10). In modular PKSs

*To whom correspondence should be addressed. Tel: +91 11 26717108; Fax: +91 11 26162125; Email: deb@nii.res.in

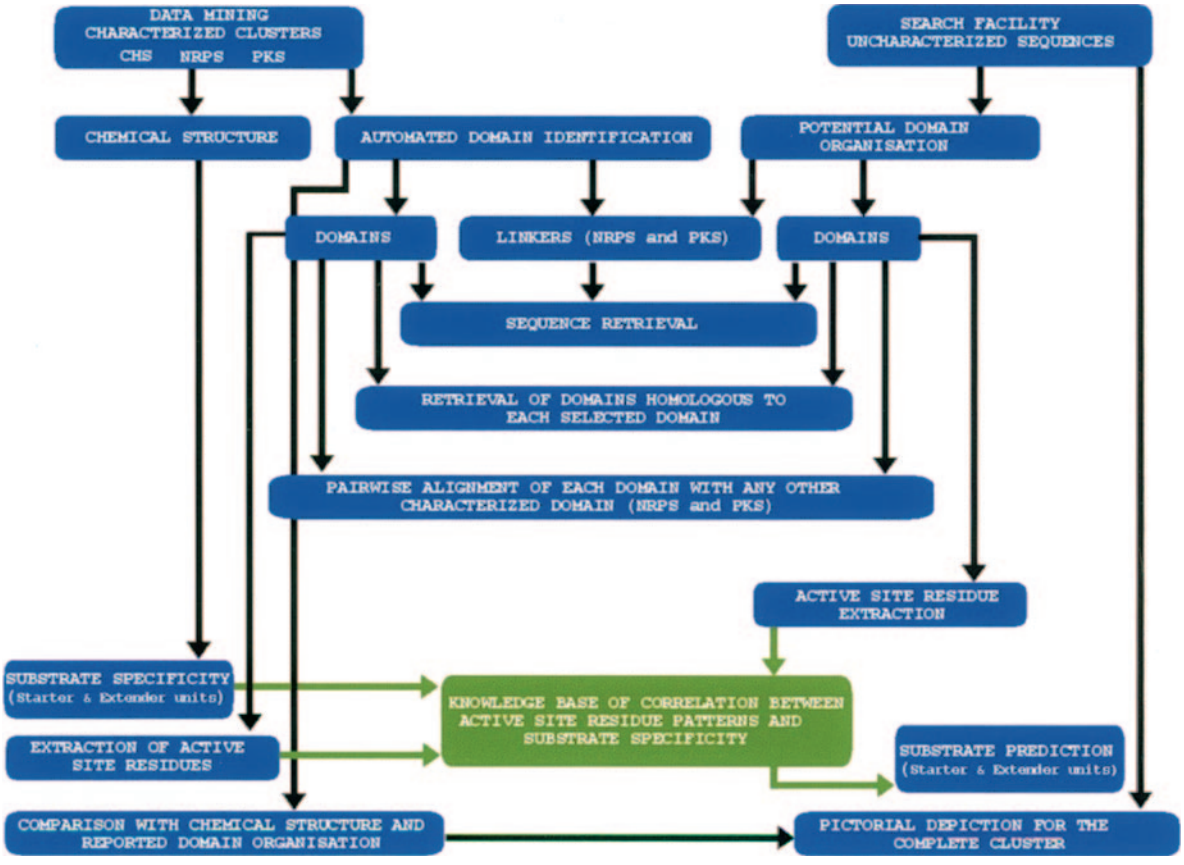

Figure 1. Flowchart depicting the organization of NRPS-PKS.

and most NRPSs, each module catalyses only one round of condensation and chain elongation/modification reaction. The number of modules in such PKSs or NRPSs correlates directly with the number of chain elongation steps during biosynthesis, and the domains present in each module dictate the chemical moiety which the given module would add to a growing chain. Experimental approaches for identification of the metabolic products of NRPS and PKS clusters require extensive bioinformatics analysis to correlate the domain organization of these gene clusters with the complex chemical structures of the metabolites. Although the sequences of various multifunctional NRPS and PKS proteins are available in different sequence databases, the organization of domains and modules, and their substrate specificities, have not been comprehensively annotated. The standard domain identification tools such as Conserved Domain Database (CDD) (11) and InterPro (12) are often found to be inadequate for accurate depiction of domain organization in these multi-functional proteins. Presently there is no resource available to predict domain organization and substrate specificities of these proteins and it is quite cumbersome to correlate these protein sequences with their metabolites. Therefore, development of automated computational tools for correct identification of NRPS/PKS domains and prediction of their substrate specificity based on identification of putative specificity-determining residues is essential for bioinformatics analyses of these proteins.

Since the various PKS domains show relatively high homology with each other within a given functional family, they can be identified by pairwise comparison with single template sequences of each domain. However, the identification of various NRPS domains with correct boundaries has been a difficult task because of the very high sequence divergence between members of a given domain family. Similarly, in contrast to the few starter/extender precursors involved in biosynthesis of polyketides, the adenylation domains of NRPSs are known to select starter/extender precursors from a pool of at least 50 different natural/unnatural amino acids. This makes the prediction of the substrate specificity of NRPS proteins a challenging task. Based on a comprehensive analysis of sequence/structural features of several experimentally characterized NRPS clusters, we have developed SEARCHNRPS, a web-based software program for prediction of the domain organization and substrate specificity of NRPS proteins. Since NRPS domains are known to occur in combination with modular PKS domains in a single open reading frame (ORF) and such hybrid NRPS/PKS clusters have been found in many microbial genomes, we have integrated SEARCHNRPS with our earlier developed tool SEARCHPKS (13,14) and made it available from a single interface named NRPS-PKS, which is capable of analysing NRPS and PKS, as well as hybrid NRPS/PKSs, sequences. Apart from modular PKSs, polyketides are also synthesized by iterative machinery (5,6) and type III polyketide synthases (15). Many complex natural products are synthesized by hybrid gene clusters which consist of NRPSs as well as various different types of PKS proteins. Therefore, we have added appropriate interfaces to NRPS-PKS for analysing iterative PKSs and CHS-like single-domain proteins belonging to the type III PKS family. In

this manuscript, we describe the method for developing NRPS-PKS, its various features for analysing NRPS/PKS gene clusters and the results of benchmarking on a test set of NRPS and PKS clusters.

METHODS

Compilation of the databases

Since NRPS-PKS uses a knowledge-based approach for prediction of domain organization and substrate specificity, a major task in development of this tool involved compilation of appropriate databases on experimentally characterized NRPS and PKS clusters. Figure 1 depicts the flowchart of the organization of these databases. The primary source of information for these databases is the protein sequences (16) of NRPS, PKS and CHS gene clusters and the chemical structures of their experimentally characterized biosynthetic products. Based on various types of bioinformatics analysis, the information on domain organization, sequences of domains and linkers, specificities of domains involved in selection of starters and extenders, and their active site residue patterns have been stored in the backend databases of NRPS-PKS. These databases are organized as four independent units named NRPSDB, PKSDB, ITERDB and CHSDB as per the type of natural products and mechanism of their biosynthesis. However, they are closely integrated with each other to permit analysis of gene clusters responsible for biosynthesis of hybrid NRPS/PKS products. Since various features of PKSDB have been reported earlier (13,14), we discuss below only the novel features of NRPSDB, ITERDB and CHSDB.

NRPSDB. NRPSDB contains information on 17 nonribosomal peptide synthetase clusters and five hybrid clusters containing NRPS as well as PKS modules. The primary task in the compilation of NRPSDB was correct identification of various NRPS domains in a polypeptide sequence. The C, A, T, epimerization (E), methyltransferase (M) and TE domains present in various NRPS clusters do not show high homology with their respective family members. Thus, unlike PKS domains, they could not be identified by using single templates. The size of C domain identified by CDD search (11) was significantly smaller than the available crystal structure of the condensation domain from vibriobactin synthetase (1L5A) (17), and in many cases the A domain depicted by CDD also differed in length from 1AMU, the crystal structure of the grsA adenylation domain (18). In order to identify the correct boundaries of C and A domains, the amino acid stretch containing the tentative C or A domains as identified by CDD and the flanking linker sequences were threaded on their structural templates using Genthreader fold recognition server (19). In the case of hybrid NRPS/PKS clusters (9), the PKS domains were identified using SEARCHPKS (14). The amino acid stretches intervening between the NRPS and PKS domains were annotated as linkers.

Domain organization in various NRPS clusters has been depicted in a pictorial format for easy correlation with the chemical structure of the NRPS product. In NRPSDB, each peptide unit in a chemical structure has been shown in a colour identical to that of the module which is responsible for its biosynthesis. By comparing the chemical structure of the NRPS product with the domain arrangement, substrate specificities have been assigned to adenylation domains responsible for selection of starter and extender precursors in a given NRPS cluster. For each adenylation domain the residues corresponding to positions 235, 236, 239, 278, 299, 301, 322, 330, 331 and 517 of 1AMU were extracted from their respective threading alignment with 1AMU (18). These 10 amino acids line the substrate binding pocket in 1 AMU. Based on earlier bioinformatics analysis (20,21), it has been proposed that corresponding residues are likely to play a key role in determining substrate specificities of other adenylation domains. Figure 2 shows a typical example of the bacitracin cluster (22) in NRPSDB. The epimerization domain (in yellow) in the fourth module of the third ORF of the bacitracin cluster is actually detected as an additional C domain by our computational method, but based on the chirality of the corresponding amino acid in the chemical structure of bacitracin it has been shown as an E domain. Images of each of the domains in Figure 2 are clickable links leading to further information. For example, on clicking the adenylation domain in first module of bacitracin, the user can get the sequence of this domain in FASTA format, its threading alignment with 1AMU and the 10 active site residues (Figure 3). Similar information on FASTA sequence, threading alignment and putative active site residues is also available for condensation domains. For each condensation domain, NRPSDB lists 10 residues corresponding to positions 125–132, 264 and 335 of 1L5A. The residues 125–131 constitute the conserved motif HHxxxDG of the C domain and its variant DxxxxD in the cyclization domain (17). The other three residues are also located in the solvent channel that is believed to be the site for the condensation reaction catalysed by this domain, and their roles in catalysis have been investigated by mutagenesis studies (17). Thus information about these 10 residues of various C domains would help in understanding whether they correlate with specificity of the condensation reaction. For the T, M and TE domains, NRPSDB only provides the sequence in FASTA format.

ITERDB. ITERDB contains information on 21 type I iterative PKSs. For each iterative PKS, ITERDB gives a depiction of the catalytic domains and chemical structure of the polyketide product. In the chemical structure, the moieties added during different iterative cycles have been depicted using different colours. Based on the chemical structures of the polyketide products, substrate specificities have been assigned to various AT domains and putative active site residues have been identified for each AT domain from their alignment with the crystal structure of acyltransferase from *Escherichia coli* FAS (1MLA) (23). ITERDB is closely integrated with PKSDB and it provides appropriate interfaces for extracting sequences of various domains and their substrate specificity and comparing them with other similar domains in ITERDB as well as PKSDB.

CHSDB. CHSDB is a database of type III polyketide synthases, which are essentially single-domain monofunctional proteins. For each protein, CHSDB provides the sequence, the chemical structures of the starter and extender substrates, information on number of condensation cycles carried out by the enzyme and the chemical structure of the enzyme-bound linear polyketide as well as the final cyclized product. Based on the crystal structures (24) of several plant

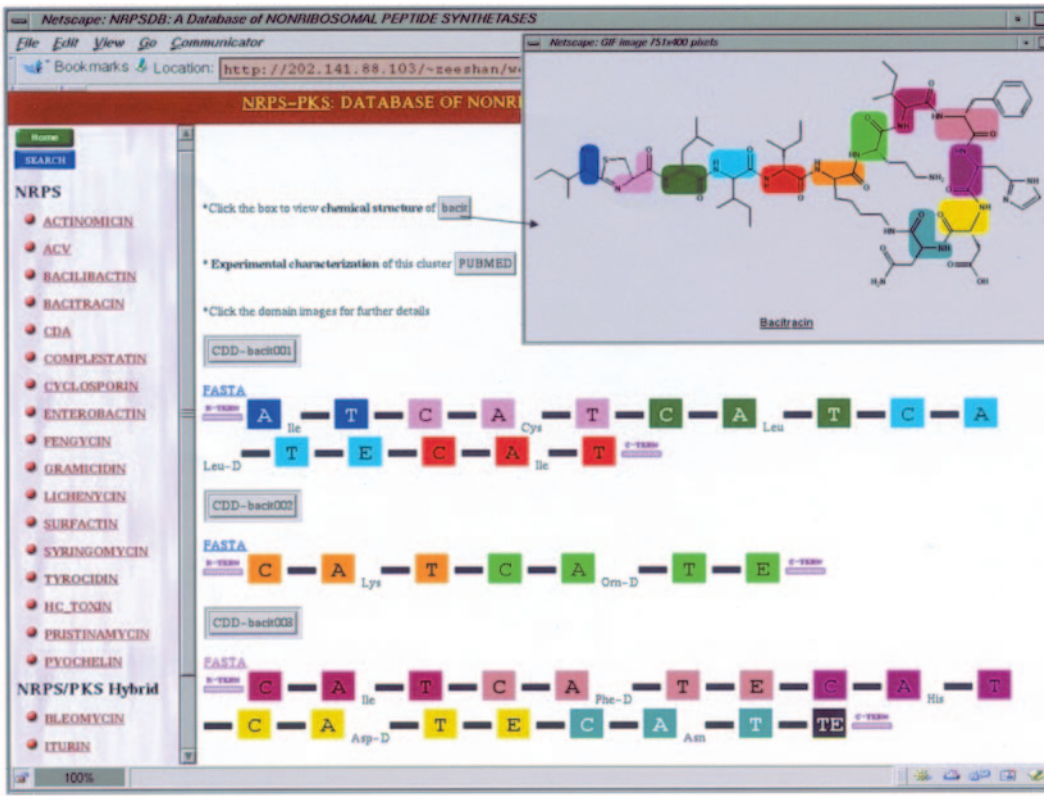

Figure 2. Depiction of the bacitracin cluster in NRPSDB. Each domain has been represented using a coloured box, the inter-domain linker regions have been depicted as filled black lines connecting the domains and the N- and C-terminal linkers have been represented as shaded lines in violet. The PUBMED button leads to the abstract of the publication describing the experimental characterization of the bacitracin cluster.

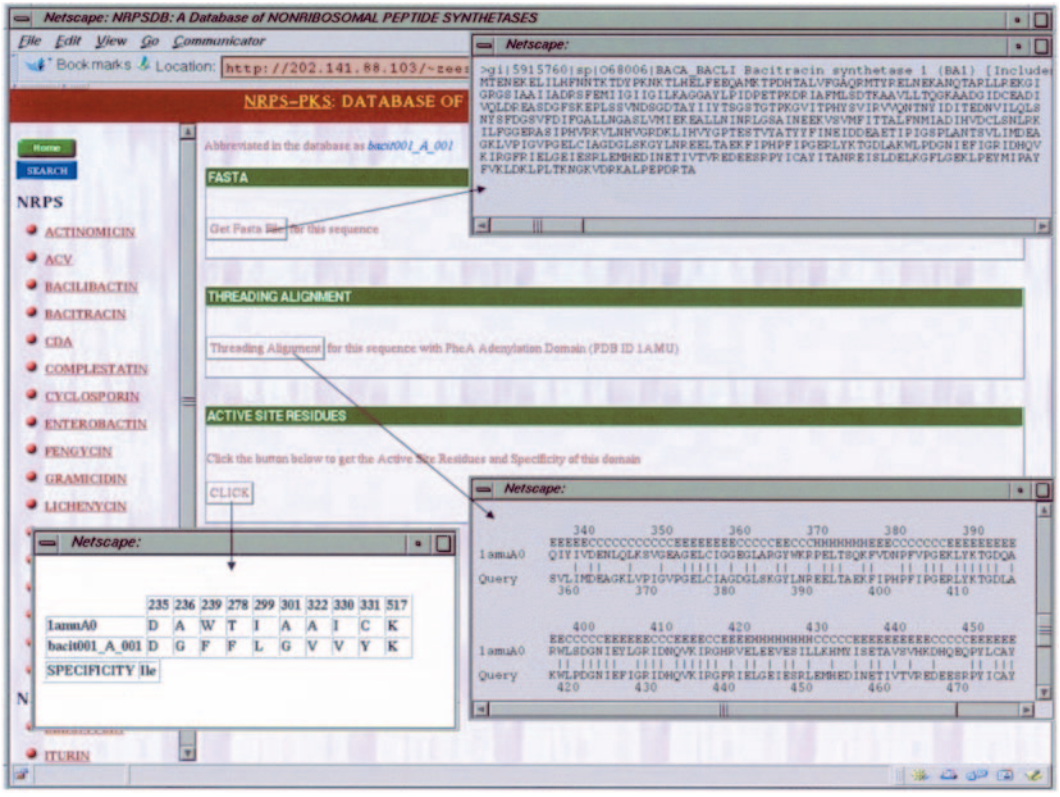

Figure 3. Screen dumps showing usage of NRPSDB for retrieval of the FASTA sequence, threading alignment and active site residues of a selected adenylation domain.

CHS proteins, it has been proposed that a set of 32 amino acids form the active site pocket and are involved in CoA binding, starter group binding, catalysis, cyclization and so on. These 32 active site residues have been extracted for each of the enzymes in CHSDB from their pairwise alignment with 1CGZ, the crystal structure of alfalfa CHS (24). CHSDB provides appropriate interfaces for comparing either the putative active site residues or the entire sequence in any selected set of CHS proteins and correlating them to their substrate specificities. Supplementary Figure 2 shows information on acridone synthase (25) as an example to highlight different features of CHSDB.

Development of the query interfaces

The NRPS and PKS clusters catalogued in NRPSDB, PKSDB, ITERDB and CHSDB have been used as the training set and the results obtained from the analysis of their sequences and structural features have been used to develop search facilities for analysing uncharacterized protein sequences containing potential NRPS, PKS or CHS domains and predicting their substrate specificities. SEARCHNRPS, the query interface for identification of NRPS domains, has been implemented using a knowledge base derived from the analysis of the NRPS domains in the 22 experimentally characterized NRPS and hybrid NRPS/PKS clusters catalogued in NRPSDB. Since NRPSDB has curated sequences of a diverse set of C, A, T, M and TE domains from 22 characterized biosynthetic clusters, for identification of a given domain all the available sequences of that domain are pairwise aligned with the query sequence using a local version of the BLAST program (downloaded from NCBI) (26). From the set of overlapping alignments, the alignment having lowest *E*-value and length above a predefined cut-off is chosen as the best match with the query. Thus, the problem associated with low homology between NRPS domains is overcome by use of multiple templates from a diverse set. Figure 4 shows a schematic diagram to depict the computational protocol used in SEARCHNRPS. It may be noted that SEARCHNRPS can also discriminate between condensation (C), cyclization (Cy) and epimerization (E) domains using information from Cy and E domains in the training set. This is achieved by aligning the query domain with the various C, Cy and E domains present in NRPSDB and identifying whether the closest match to the query domain is a condensation, cyclization or epimerization domain. Prediction of a Cy domain is further confirmed by searching for the conserved DxxxxD motif. This approach for discriminating between C, Cy and E domains has been tested by carrying out benchmarking on a test set. Thus SEARCHNRPS can correctly predict C, Cy, E, A, M and T domains in a query sequence. Other domains which occur in NRPS clusters with very low frequency, e.g. oxidation, reductase and formylation, cannot be predicted without significantly enlarging our training set of gene clusters. Thus, the present version of NRPS-PKS would predict only the C, Cy, E, A, M and T domains, and any other domain would be annotated as a linker. However, our sequence analysis has indicated that the linker regions between these domains are small in length. Thus, detection of any long inter-domain NRPS protein sequences would suggest the possible presence of additional domains. SEARCHNRPS depicts such regions in red so that those regions can be analysed in detail by other programs for the presence of additional NRPS domains.

Apart from deciphering the domain organization, SEARCHNRPS also identifies the putative active site residues of A and C domains present in the query from their alignment with structural templates. By comparing these putative active site residues with the active sites of A domains of known specificity in NRPSDB, SEARCHNRPS attempts to predict the specificity of the A domain in the query. It may be noted that there is one more web server (http://raynam.chm.jhu.edu/~nrps/) which predicts specificity of adenylation domains using a similar knowledge-based approach (21). SEARCHNRPS also has options for extracting the alignment of a selected NRPS domain with homologous domains present in NRPSDB. The interfaces for analysis of iterative PKSs and type III PKSs have been developed using protocols similar to those of SEARCHPKS, the query interface for modular PKSs. Since the backend databases are integrated with each other, NRPS-PKS can analyse hybrid NRPS/PKS clusters in a query by combination of SEARCHNRPS and SEARCHPKS. The query interfaces of NRPSDB, PKSDB, ITERDB and CHSDB have options for aligning the sequence of a query domain with homologous structures in PDB, and from the alignment page links are also provided to the corresponding structure entry at the RCSB site. This structure link option is available for KS, AT, ACP and TE domains of PKSs, and C, Cy, E, A, T and TE domains of NRPSs, because structural templates for these domains could be identified with certainty.

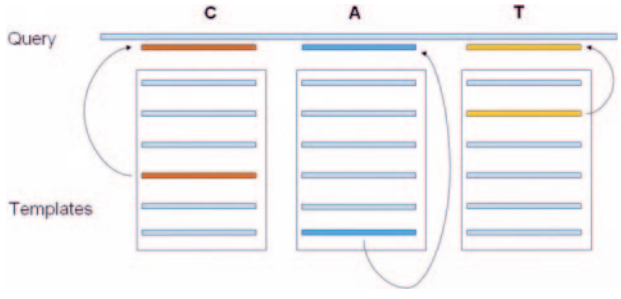

Figure 4. Schematic diagram showing the multiple template method for detection of NRPS domains. For each domain type a diverse set of sequences from the training set are used as templates. Domains in the query are identified by simple BLAST using these multiple templates and picking the best match.

RESULTS

In order to benchmark the prediction accuracy of NRPS-PKS, we have used a set of experimentally characterized PKS, NRPS and hybrid NRPS/PKS clusters which are not included in the backend databases. Since the biosynthetic product is known for gene clusters in this test set, the correct domain organization and substrate specificity can be inferred from the chemical structure of the product metabolite and can be compared with the *in silico* predictions of NRPS-PKS.

Figures 5 and 6 show a typical example of the usage of NRPS-PKS for depicting domain organization in a query sequence from a hybrid NRPS-PKS cluster. When the amino acid sequence of the query protein has been submitted in the text box on the search page, the program gives a pictorial

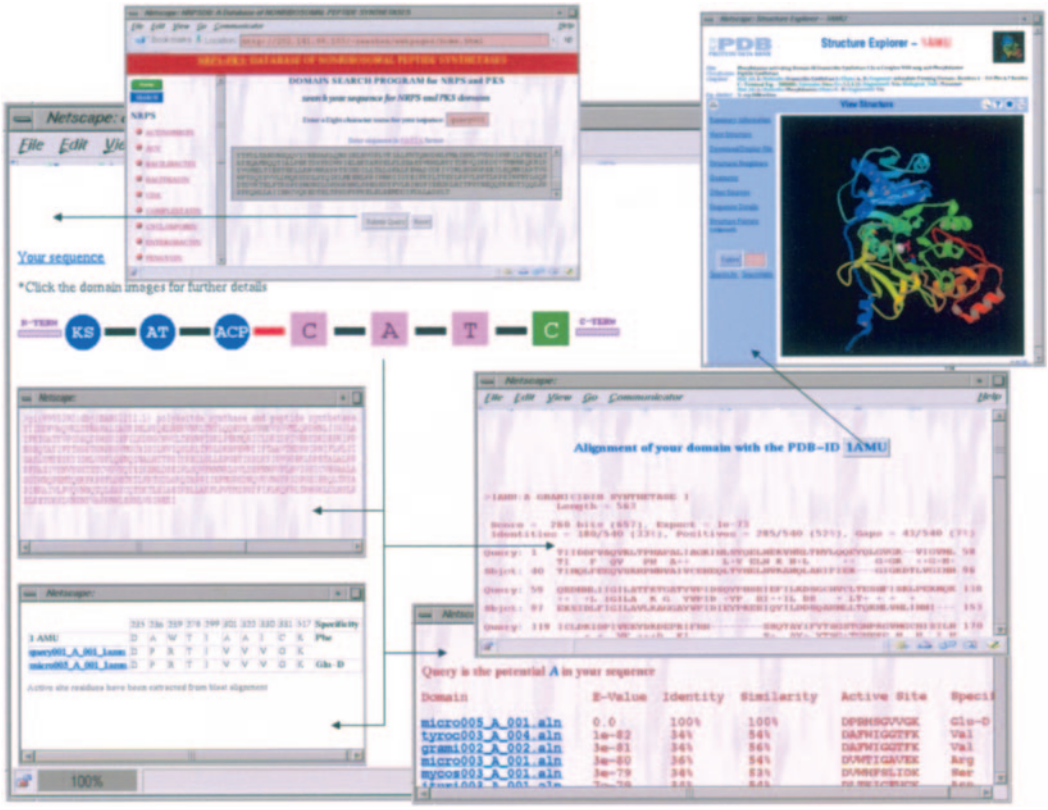

Figure 5. Typical use of the query interface of NRPSDB for analysis of a hybrid NRPS/PKS cluster. Upon submission of a query sequence, the program gives a pictorial depiction of domain organization similar to those for characterized clusters in NRPS-PKS. The red-coloured linker between PKS and NRPS modules indicates the possible presence of domains which could not be detected by NRPS-PKS. Clicking on the image of a domain leads to further details. On selecting an adenylation domain in the query, the user can get its sequence, active site residues, predicted specificity, list of homologous domains along with their degree of similarity and specificity, and also information on homologous structures from PDB.

depiction of domain organization. On clicking the images of NRPS domains, the user is led to the screens containing FASTA sequence, alignment with homologous structures in PDB and links to those PDB entries, comparison with similar domains in NRPSDB and also putative active site residues of C and A domains. As can be seen, apart from the FASTA sequence of the depicted domains and their alignment with homologous sequences and structures, the program also predicts substrate specificity based on the A domain of known specificity that has the closest active site match to the query. Links are also given to the alignments from which active site residues have been extracted. Similar analysis is also possible for PKS domains since NRPS-PKS uses SEARCHPKS as query interface. For example, the AT domain shown in Figures 5 and 6, can be compared with sequences of homologous domains in PKSDB and ITERDB, and also with homologous structures in PDB. Figure 7 shows typical screen shots of various features for analysis of PKS domains. The detailed prediction results for all the NRPS, PKS and hybrid NRPS/PKS clusters in the test set are available at http://www.nii.res.in/nrps-pks/benchmark.html. In Supplementary Table 1, we report a summary of the prediction results. As can be seen, NRPS-PKS is able to detect all the A domains, 106 out of 108 C/Cy/E domains and 105 out of 110 T domains, as well as most of the M and TE domains. In view of the extremely low sequence similarity between C domains, the high prediction accuracy of our multiple template approach is encouraging. It may be noted that profile-based domain identification tools such as CDD (11) and InterPro (12) fail to identify the correct domain boundaries of C domains. Our domain identification protocol has been successful because the correct boundaries of the template C domains have been detected by threading analysis and the template set adequately represents the sequence diversity seen in condensation domains. NRPS-PKS is also able to predict the exact substrate specificity for 52 A domains, and a chemically similar amino acid is predicted for another 16 A domains. There are nine A domains in the test set whose substrates are not represented in the training set; thus their specificities are unlikely to be predicted by our knowledge-based approach. Therefore, the accuracy of NRPS-PKS for prediction of substrates is 85%. Considering the large variety of substrates the adenylation domains can accept, even at 85% accuracy for substrate prediction, NRPS-PKS would be a valuable tool for experimental characterization of metabolic products. It is also clear from our analysis that with an increase in the data set for representative substrates, the prediction accuracy can be improved further. In the case of PKS domains (Supplementary Table 1B) NRPS-PKS is also able to correctly identify all but one KS domain, 34 out of 38 dehydratase (DH) domains and all the AT, enoyl-reductase (ER), ketoreductase (KR) and ACP domains. Even though NRPS-PKS detects all the ACP domains in the test set, it also gives false positive prediction of six ACP domains, which can be excluded by analysis of conserved signature

Figure 6. Screen shot depicting usage of the query interface of NRPS-PKS for extracting the active site residues of a selected condensation domain and obtaining its pair alignment with homologous condensation domains in NRPSDB. For each homologous C domain, the program also lists the amino acid pair it is condensing, and whether it is involved in a cyclization or epimerization reaction.

motif for ACP domains. Out of the 56 AT domains in the test set, NRPS-PKS can correctly predict specificity for 52 AT domains, thus giving a prediction accuracy of 93%. The results of our benchmarking on 32 NRPS, PKS and hybrid NRPS/PKS clusters indicate that NRPS-PKS can predict domain organization and substrate specificity with a very high accuracy; thus it can be a powerful tool for analysing uncharacterized NRPS and PKS clusters.

Similar analysis for CHS-like type III PKS proteins indicates that, based on the putative active site residues, NRPS-PKS can predict substrate specificity of type III PKS proteins found in plant and bacterial genomes. In a recent work, interesting clues about substrate specificity of pks18, a type III PKS protein from *Mycobacterium tuberculosis*, were obtained from the analysis of its active site residues using CHSDB and its preference for an unusual long chain aliphatic starter unit was demonstrated experimentally (27).

Apart from analysing uncharacterized NRPS and PKS clusters, NRPS-PKS can also be used for comparative analysis of various experimentally characterized clusters catalogued in backend databases as well as in the test sets. Since putative active site residues and substrate specificities are available for a large number of CHS-like proteins, acyltransferase and adenylation domains, it can help in designing site directed mutagenesis experiments for making altered natural products. Similarly, information on linker sequences and tools to search for homologous domains with desired specificity can help in domain-swapping experiments (28).

DISCUSSION

We have organized the sequence information on various experimentally characterized NRPS and PKS gene clusters in the form of searchable computerized databases. These gene clusters are used as a training set, and based on the analysis of their sequence and structural features, we have developed NRPS-PKS, a knowledge-based tool for analysing putative NRPS and PKS gene clusters. NRPS-PKS facilitates easy extraction from a polypeptide sequence of various domains and identification of their catalytic activity, active site residues, substrate specificity and so on, and it permits comparison of NRPS/PKS domains in terms of their sequence similarity, substrate specificity and active site motifs. Benchmarking on a test set of NRPS and PKS clusters has established that it can correctly identify domain boundaries and predict specificity for starter/extender precursor units with a very high accuracy. To the best of our knowledge, NRPS/PKS is the only available tool which can correctly identify various NRPS and PKS domains and predict their specificities. This makes NRPS-PKS a valuable resource for providing leads to decipher the natural products biosynthesized by NRPS/PKS clusters found in newly sequenced microbial genomes. The training and test sets of NRPS-PKS have information on 307 ORFs, 2223 functional protein domains, 68 starter/extender precursors and their specific recognition motifs, and also the chemical structure of 101 natural products from four different families. Therefore, NRPS-PKS is also a powerful tool for

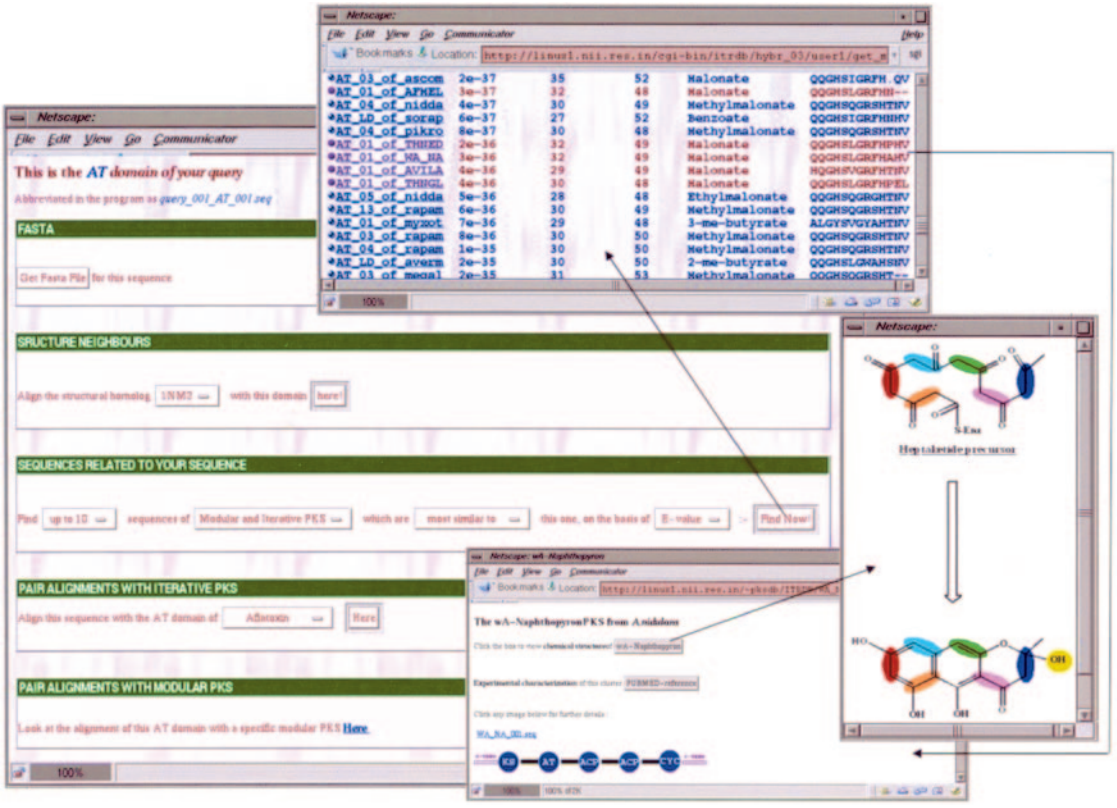

Figure 7. Features of NRPS-PKS for analysis of PKS domains in the hybrid NRPS/PKS query. The software provides interfaces for extracting its sequence in FASTA format, its pair alignment with any other AT domain in PKSDB or ITERDB, alignment with homologous structures in PDB and a list of homologous AT domains from ITERDB and PKSDB along with their active site residues. Iterative AT matches are marked with a pink ball and modular ones with a blue ball. From the alignment with homologous domains, the program also provides link to the depiction of the PKS cluster containing the homologous domain.

designing experiments to engineer 'unnatural' natural products based on analysis of known gene clusters.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Dr Sandip K. Basu for his encouragement and support. M.Z.A. and G.Y. are recipients of Senior Research Fellowships from CSIR, India. R.S.G. is a Wellcome Trust International Senior Research Fellow for biomedical sciences in India. The work has been supported by grants to the National Institute of Immunology from the Department of Biotechnology, Government of India. Computational resources provided under the BTIS project of DBT, India are gratefully acknowledged.

REFERENCES

1. Cane,D.E, Walsh,C.T. and Khosla,C. (1998) Harnessing the biosynthetic code: combinations, permutations, and mutations. *Science*, **282**, 63–68.
2. Cane,D.E. and Walsh,C.T. (1999) The parallel and convergent universes of polyketide synthases and nonribosomal peptide synthetases. *Chem. Biol*., **6**, R319–R325.
3. Marahiel,M.A., Stachelhaus,T., and Mootz,H.D. (1997) Modular peptide synthetases involved in non-ribosomal peptide synthesis. *Chem. Rev*., **97**, 2615–2673.
4. Mootz,H.D., Schwarzer,D. and Marahiel,M.A. (2002) Ways of assembling complex natural products on modular nonribosomal peptide synthetases. *ChemBioChem*., **3**, 490–504.
5. Gokhale,R.S. and Tuteja,D. (2001) Biochemistry of polyketide synthases. In Rehm,H.J. (ed.), *Biotechnology*, 2nd ed. WILEY-VCH, Weinheim, Vol. 10, pp. 341–372.
6. Hopwood,D.A. (1997) Genetic contributions to understanding polyketide synthases. *Chem. Rev*., **97**, 2465–2498.
7. Staunton,J. and Weissman,K.J. (2001) Polyketide biosynthesis: a millennium review. *Nat. Prod. Rep*., **18**, 380–416.
8. Khosla,C., Gokhale,R.S., Jacobsen,J.R. and Cane,D.E. (1999) Tolerance and specificity of polyketide synthases. *Annu. Rev. Biochem*., **68**, 219–253.
9. Du,L. and Shen,B. (2001) Biosynthesis of hybrid peptide–polyketide natural products. *Curr. Opin. Drug Discov. Devel*., **4**, 215–228.
10. Gokhale,R.S. and Khosla,C. (2000) Role of linkers in communication between protein modules. *Curr. Opin. Chem. Biol*., **4**, 22–27.
11. Marchler-Bauer,A., Panchenko,A.R., Shoemaker,B.A., Thiessen, P.A., Geer, L.Y. and Bryant, S.H. (2002) CDD: a database of conserved domain alignments with links to domain three-dimensional structure. *Nucleic Acids Res*., **30**, 281–283.
12. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Barrell,D., Bateman,A., Binns,D., Biswas,M., Bradley,P., Bork,P. *et al*. (2003) The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res*., **31**, 315–318.
13. Yadav,G., Gokhale,R.S. and Mohanty,D. (2003) Computational approach for prediction of domain organization and substrate specificity of modular polyketide synthases. *J. Mol. Biol*., **328**, 335–363.

14. Yadav,G., Gokhale,R.S. and Mohanty,D. (2003) SEARCHPKS: a program for detection and analysis of polyketide synthase domains. *Nucleic Acids Res.*, **31**, 3654–3658.
15. Austin,M.B. and Noel,J.P. (2003) The Chalcone synthase superfamily of type III polyketide synthases. *Nat. Prod. Rep.*, **20**, 79–110.
16. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2003) GenBank. *Nucleic Acids Res.*, **31**, 23–27.
17. Keating,T.A., Marshall,C.G., Walsh,C.T. and Keating,A.E. (2002) The structure of VibH represents nonribosomal peptide synthetase condensation, cyclization and epimerization domains. *Nat. Struct. Biol.*, **9**, 522–526.
18. Conti,E., Stachelhaus,T., Marahiel,M.A. and Brick,P. (1997) Structural basis for the activation of phenylalanine in the nonribosomal biosynthesis of gramicidin S. *EMBO J.*, **16**, 4174–4183.
19. Jones,D.T. (1999) GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *J. Mol. Biol.*, **287**, 797–815.
20. Stachelhaus,T., Mootz,H.D. and Marahiel,M.A. (1999) The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases. *Chem. Biol.*, **6**, 493–505.
21. Challis,G.L., Ravel,J. and Townsend,C.A. (2000) Predictive, structure-based model of amino acid recognition by nonribosomal peptide synthetase adenylation domains. *Chem. Biol.*, **7**, 211–224.
22. Konz,D., Klens,A., Schorgendorfer,K. and Marahiel,M.A. (1997) The bacitracin biosynthesis operon of *Bacillus licheniformis* ATCC 10716: molecular characterization of three multi-modular peptide synthetases. *Chem. Biol.*, **4**, 927–937.
23. Serre,L., Verbree,E.C., Dauter,Z., Stuitje,A.R. and Derewenda,Z.S. (1995) The *E.coli* malonyl CoA: acyl carrier protein transacylase at 1.5 Å resolution. Crystal structure of a FAS component. *J. Biol. Chem.*, **270**, 12961–12964.
24. Ferrer,J.L., Jez,J.M., Bowman,M.E., Dixon,R.A. and Noel,J.P. (1999) Structure of chalcone synthase and the molecular basis of plant polyketide biosynthesis. *Nat. Struct. Biol.*, **6**, 775–784.
25. JunghannsK.T., Kneusel,R.E., Baumert,A., Maier,W., Groger,D. and Matern,U. (1995) Molecular cloning and heterologous expression of acridone synthase from elicited *Ruta graveolens* L. cell suspension cultures. *Plant Mol. Biol.*, **27**, 681–692.
26. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
27. Saxena,P., Yadav,G., Mohanty,D. and Gokhale,R.S. (2003) A new family of type III polyketide synthases in *Mycobacterium tuberculosis*. *J. Biol. Chem.*, **278**, 44780–44790.
28. Katz,L. and McDaniel,R. (1999) Novel macrolides through genetic engineering. *Med. Res. Rev.*, **19**, 543–558.