J Ind Microbiol Biotechnol (2013) 40:653–659
DOI 10.1007/s10295-013-1252-z

NATURAL PRODUCTS

Databases of the thiotemplate modular systems (*CSDB*) and their in silico recombinants (*r-CSDB*)

Janko Diminic · Jurica Zucko · Ida Trninic Ruzic · Ranko Gacesa · Daslav Hranueli · Paul F. Long · John Cullum · Antonio Starcevic

Received: 19 December 2012 / Accepted: 22 February 2013 / Published online: 16 March 2013

Abstract Modular biosynthetic clusters are responsible for the synthesis of many important pharmaceutical products. They include polyketide synthases (PKS clusters), non-ribosomal synthetases (NRPS clusters), and mixed clusters (containing both PKS and NRPS modules). The *ClustScan* database (*CSDB*) contains highly annotated descriptions of 170 clusters. The database has a hierarchical organization, which allows easy extraction of DNA and protein sequences of polypeptides, modules, and domains as well as an organization of the annotation so as to be able to predict the product chemistry to view it or export it in a standard SMILES format. The recombinant *ClustScan* database contains information about predicted recombinants between PKS clusters. The recombinants are generated by modeling homologous recombination and are associated with annotation and prediction of product chemistry automatically generated by the model. The database contains over 20,000 recombinants and is a resource for in silico approaches to detecting promising new compounds. Methods are available to construct the corresponding recombinants in the laboratory.

Keywords Polyketides · Non-ribosomal peptides · PKS/NRPS hybrids · Computer programs · Databases

Electronic supplementary material The online version of this article (doi:10.1007/s10295-013-1252-z) contains supplementary material, which is available to authorized users.

J. Diminic · J. Zucko · I. T. Ruzic · 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

Present Address:
I. T. Ruzic
Croatian Accreditation Agency, Vukovarska 78, 10000 Zagreb, Croatia

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

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

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

Introduction

Modular polyketide synthase (PKS), non-ribosomal peptide synthetase (NRPS), and polyketide/peptide (PKS/NRPS) hybrid gene clusters, which are collectively called thiotemplate modular systems (TMS) [7], are gene clusters whose protein products are involved in the biosynthesis of very important classes of compounds that have many useful biological activities [6, 9, 13, 32]. These modular biosynthetic clusters are particularly interesting as they function according to a "building block" principle in which each module is usually responsible for a single extension step during the synthesis of the product. DNA sequencing of PKS, NRPS, and hybrid gene clusters showed that their products are multi-functional enzymes with multi-modular organization. The modules themselves are made up of several domains. A PKS extension module requires the three domains: ketosynthase (KS), acyltransferase (AT), and acyl carrier protein (ACP), but often contains further reduction domains, which modify the incorporated substrate. Similarly, an NRPS extension

module requires the three domains: condensation (C), adenylation (A) and peptidyl carrier protein (PCP) and sometimes contains further domains to modify the substrate. The final module of a gene cluster usually ends with a thioesterase (TE) domain, which is responsible for the detachment of the product from the enzyme and its cyclization. After that, post-polyketide or post-peptide "decorating" enzymes produce the final structure [for reviews see: 8, 12, 15, 21, 28]. The modular organization allows much of the chemical structure of a product to be predicted from the DNA sequence [32].

The DNA sequences of many clusters are available in public databases, but these are usually not annotated in detail; in many cases coding regions are annotated as polypeptides containing PKS or NRPS modules, but no further details of module specificities are given. There are also many chemical structures of products in databases, but they are not usually linked to the biosynthetic clusters. A number of computer programs have been developed for the analysis of PKS, NRPS and hybrid gene clusters: *SEARCHPKS* [29], *DecipherIT* [30], *NRPSpredictor* 1 and 2 [19, 20], *Biogenerator* [31], *MAPSI*, [25], *ClustScan* [22], CLUSEAN [26], NP.searcher [16], *SBSPKS* [1], and *antiSMASH* [17]. The de novo analysis of new sequences is relatively time-consuming and generates considerable data about the specificities of domains and modules and the chemical structures of the products. It is, therefore, attractive to incorporate the results of the analyses in a database, which allows easy visualization of the biosynthetic steps and extraction of DNA and protein sequences of domains.

The PKSDB-NRPSDB database [1, 2], which is associated with the *SEARCHPKS* analysis program, holds data on publicly available polyketide, peptide, and hybrid gene clusters including domain and module architecture and the chemical structures of the gene cluster products. Another useful publicly available polyketide database is the ASMPKS database [25], which was developed on the basis of the *MAPSI* program. The DoBISCUIT database holds data on PKS and NRPS clusters derived from the literature with a manual re-annotation of the clusters to achieve uniform descriptions of the modules and domains [11]. The Norine database [5] contains information about the chemistry of non-ribosomal peptides, but does not contain information on DNA sequences. All these databases rely on extensive manual curation, so that they require considerable effort to add new clusters and there are dangers of errors being introduced. Recently, ClusterMine360, a database of microbial PKS/NRPS biosynthesis was published [4; http://clustermine360.ca/Default.aspx], which attempts to overcome the labor of the curation problem by public sourcing of entries.

The problem of populating databases is becoming ever greater with the rapid progress in sequencing technology. For example, even using a conservative estimate that every actinobacterial genome contains ten TMS gene clusters and that 1,000 sequenced genomes will soon be available, there will soon be 10,000 new TMS gene clusters, potentially encoding novel chemical entities [22, 32]. The *ClustScan* program [22] was developed to analyze modular clusters, but unlike the other programs, takes a "top-down" approach to the annotation of gene clusters so that the cluster is also considered as a whole unit. A special XML data structure was developed so that the polypeptides, modules, and domains are organized in a hierarchical way and it is possible to predict the structures of the products. In this paper, we describe how this data structure allows us to construct the *ClustScan* database (*CSDB*), which can be populated directly from the *ClustScan* program without any further manual curation steps.

The *CompGen* program [23] models homologous recombination between modular PKS clusters and was developed to help overcome problems of low yield in genetically manipulated clusters, when inappropriate recombination junctions are employed. It utilizes the XML data structure from *ClustScan* to predict the chemical structures of products from recombinant clusters. Continuing advances in computer-aided drug design technology [14] are making it possible to predict which of these chemical entities are likely to possess useful biological activities. Initial studies with 47 cluster pairs generated nearly 12,000 polyketide structures [24] so that, if 1,000 gene cluster sequences were available, in silico recombination should generate over 5,000,000 polyketide structures, mostly novel chemical entities. Most exciting of all, when such a product looks promising in silico, a "designer bug" can be created in the laboratory to produce it [23, 24, 32]. In this paper, we describe the recombinant *ClustScan* database (*rCSDB*), which contains information about in silico recombinants.

Materials and methods

The databases are implemented using MySQL [http://www.mysql.com/]. The Web interfaces were written using Google Web Toolkit (GWT; https://developers.google.com/web-toolkit/overview) technology with the Apache tomcat server [http://tomcat.apache.org/]. The DNA sequences of the clusters were obtained from GenBank (http://www.ncbi.nlm.nih.gov/) and were annotated using *ClustScan* [22]. The in silico recombinants were obtained using the *CompGen* program [23, 24]. The *CSDB* and *r-CSDB* databases are freely available at http://bioserv.pbf.hr/cms/.

Results

The ClusScan database (CSDB)

Each cluster has a hierarchical structure with genes containing modules that are made up of domains. This is reflected in the organization of the database (Fig. 1); a detailed description of the structure of tables is shown in Fig. 1S in the supplementary material. The domains are classified into types based on activities and specificities. The domain properties determine the nature of the chemical extender unit, which is built in by the module and the chemical structure of the extender unit (in an isomeric SMILES format, [27]) is associated with each module. All of these data are generated by a *ClustScan* analysis of the cluster DNA sequence. If literature references for the cluster are available, they can be added manually to the cluster entry. The information from *ClustScan* can automatically generate the chemical structure of the linear product, but automatic prediction of the cyclization, which occurs for most products, is not at present feasible. However, if the cyclized structure is known, it can be added to the aglycone field of the entry. At present, there are 57 PKS, 51 NRPS, and 62 hybrid gene clusters in *CSDB* (170 in total); the clusters are listed and can be selected by clicking (Fig. 2S of the supplementary material). These clusters are derived from actinomycetes, myxobacteria, and *Bacillus* species, but the database structure is suitable for any bacterial modular clusters.

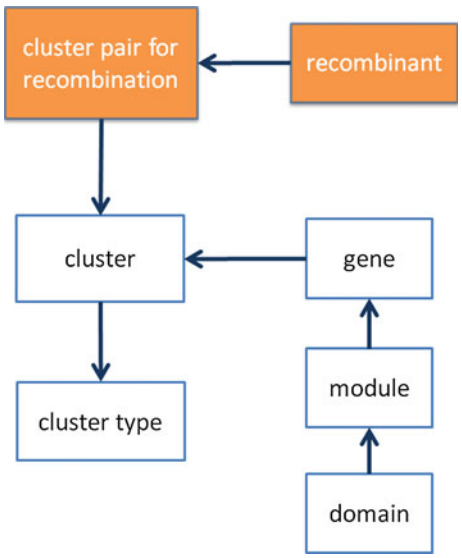

Fig. 1 Structure of *CSDB* and *r-CSDB*. Each *block* represents a table in MySQL and details of the tables are shown in Fig. 1S. The *r-CSDB*-specific tables are in *orange*. *r-CSDB* uses information from *CSDB* to reconstruct the sequence and genetic structure of recombinants

The specially written Web interface allows several clusters to be opened in separate tabs facilitating comparison. The genetic structure of each cluster is represented in a cartoon (Fig. 2a) with genes, modules, and domains. The cartoon image can be downloaded in PNG-format (Fig. 2b). The details window below (Fig. 2c) shows general information about the cluster and its product and literature references, if available in the database. It also shows the linear chemical structure and (if available) the cyclized aglycone structure. Clicking on a module changes the details window to module information including location in the DNA and the nature of the extender unit incorporated. Similarly, clicking on a domain gives information of the location, activity, and specificity of a domain. It is possible to download DNA or protein sequences of the genes, modules, or domains as well as the DNA sequence of the whole cluster (see: Figs. 3S to 6S in the supplementary material).

There is a search function that allows the user to find clusters, modules, and domains with particular properties. For instance, it is possible to list all AT domains that utilize a particular substrate (see Fig. 7S in the supplementary material). The DNA or protein sequences of chosen domains can be downloaded for analysis. This facility to obtain collections of modules or domain solutions from *CSDB* has been used to generate data used in several publications. Phylogenetic analyses using KS and C domains, respectively, were used to help assemble PKS and NRPS clusters in the genome sequence of *Streptomyces tsukubaensis* [3]. PKS modules extracted from *CSDB* were used for an analysis of synonymous vs. non-synonymous codons in order to identify regions undergoing strong selection, which helps in the selection of targets for in vitro manipulation of clusters [33]. Clusters and domains were used for the analysis of PKS cluster evolution, which identified gene conversion and horizontal gene transfer as important forces [34].

The recombinant-ClustScan database (*r-CSDB*)

r-CSDB uses *CSDB* for information on parent clusters. It has tables for each pair of clusters that have been used for recombination, which are associated with tables corresponding to each recombinant, which detail the recombination sites in each recombinant (Fig. 1, Fig. 1S in the supplementary material). The DNA sequences of the recombinants and the genetic structure of the recombinant clusters are generated from this information using data from *CSDB*. At present, there are 47 PKS parental clusters and 20,187 recombinant gene clusters in the *r-CSDB* database, which generate 11,796 unique compounds (see Fig. 8S in the supplementary material). Data can be entered into the database by importing further parental clusters

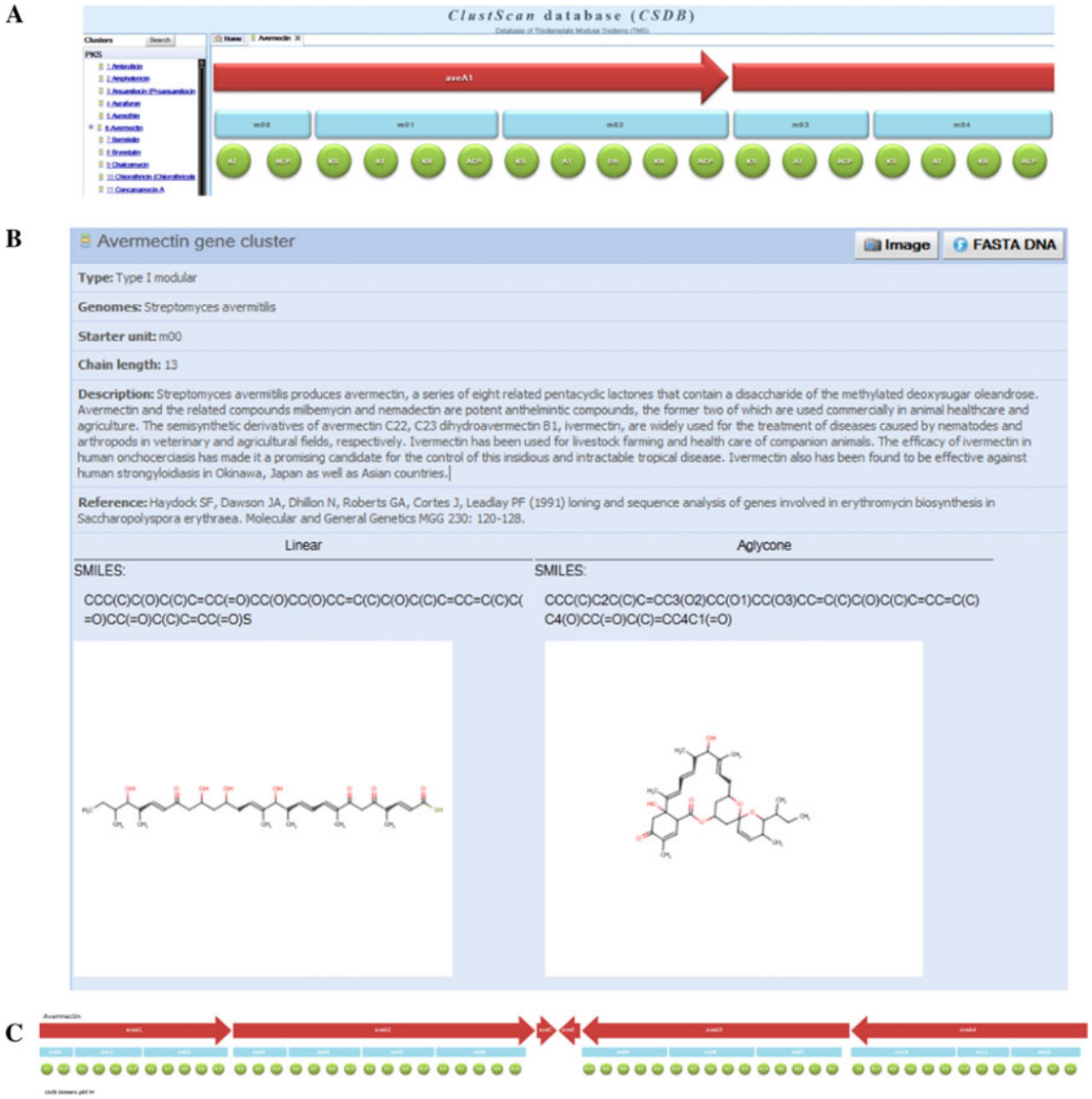

Fig. 2 Screenshots of the avermectin cluster in *CSDB*. **a** Cartoon view of the genetic structure of the cluster: genes are shown as *red arrows*, modules as *blue bars*, and domains as *green circles*. **b** General information about the cluster with chemical structures of the linear backbone and aglycone together with the isomeric SMILES descriptions. Clicking on the "Image" button produces a window containing the cartoon of cluster genetic structure in PNG format (**c**)

from the *CSDB* database and data about recombination sites from the *CompGen* program allowing easy expansion of the database.

r-CSDB can be accessed using a specially written Web interface that generates the information about recombinant DNA sequences and the biosynthetic pathways and products from the database entry. After selecting one parental cluster, a list of clusters for which there are recombinants is shown. Selection of the second parent shows a list of recombinants and by clicking on the list it is possible to see the details (Fig. 3). The locations of the recombination sites are shown and a cartoon showing the genetic structures of the parents and the recombinant is shown. The predicted chemical structure of the recombinant product is shown as an isomeric SMILES and as a chemical structure. It is possible to download the DNA sequence of the recombinant. It is also possible to get an overview of the chemical properties of the recombinants generated from a particular parent. For example, Fig. 4 shows the distribution of molecular weights and degree of reduction for recombinants generated between the avermectin cluster and all other clusters. It is also possible to display the information for recombinants between a single pair of clusters.

The Web interface offers a simple and convenient tool for working with small numbers of recombinants. When

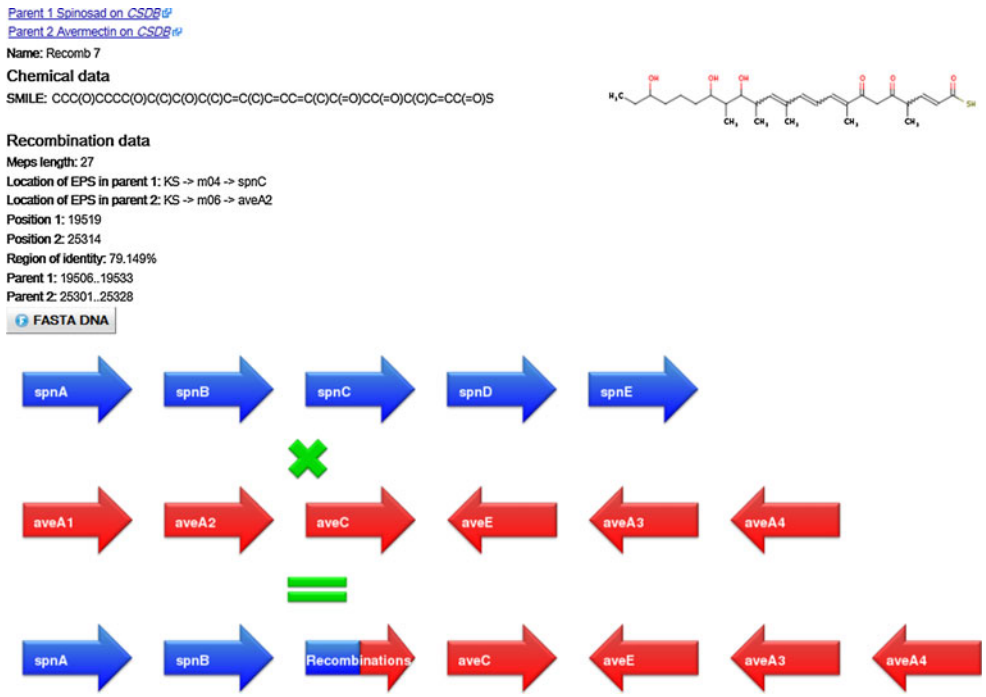

Fig. 3 Screenshot of *r-CSDB* information about a recombinant between the avermectin and spinosad clusters. Information about the locations of the recombination sites in the two clusters and the predicted chemical structure of the product as well as the SMILES description are shown. The cartoon shows the genetic structures of the parents and the recombinant. By clicking on the FASTA button the DNA sequence of the recombinant can be downloaded

larger quantities of data are needed (e.g., for processing a large number of recombinant chemical structures with chemoinformatic programs) the standard tools incorporated in MySQL can be used to export the appropriate information.

Discussion

The *CSDB* and *r-CSDB* databases are based on the XML data structure developed for *ClustScan* [22] and are implemented using MySQL. This links the DNA sequences to the proteins and the chemical structure of the products. This allows easy direct entry of new data, avoiding errors associated with manual steps and considerably reducing the labor of adding new clusters, which is a major factor reducing the growth of most other databases for modular biosynthetic clusters [1, 2, 5, 25]. ClusterMine360 [4] tries to overcome this problem by public sourcing of entries, but it is not yet clear how successful this is for accuracy and growth of the database. The data structure of *CSDB* is also well suited in dealing with the consequences of genetic manipulation of the clusters. The present data structure is designed for bacterial clusters and it is easy to add profiles allowing the identification of novel domains or domains in unusual organisms, which may not be recognized by the standard profiles. At present, the data structure is not suitable for clusters derived from eukaryotes (e.g., fungi, slime moulds, plants), because it does not allow gene models with introns. If such models were implemented, it would be possible to include such clusters. However, most of the known eukaryotic clusters involve either iterative PKSs, where the chemical structure cannot be predicted as the number of iterative steps is unknown, or clusters for β-lactam synthesis, where the conserved NRPS synthesizes the standard tripeptide precursor, i.e., the structure is clear without further analysis. Other databases allow the extraction of sequences associated with particular clusters. However, *CSDB* also allows the extraction of sets of sequences with particular properties (e.g., AT domains with the same substrate specificity), which is a very useful resource for a variety of purposes including manipulation of clusters and evolutionary studies. Evolutionary studies are of more than just academic interest as they can help to assemble clusters in genome sequencing projects [3] or suggest strategies to develop new recombinant products [33, 34].

The *r-CSDB* is a unique database of in silico recombinants, which will become of ever-increasing interest. On the one hand, computer-aided drug design technology is advancing rapidly [14], which increases the efficiency of recognizing promising candidates. On the other hand, advances in cloning technology and synthetic biology make it easier to construct recombinants in the laboratory to produce the desired compounds. Vectors and tools are available to construct recombinants using traditional genetic

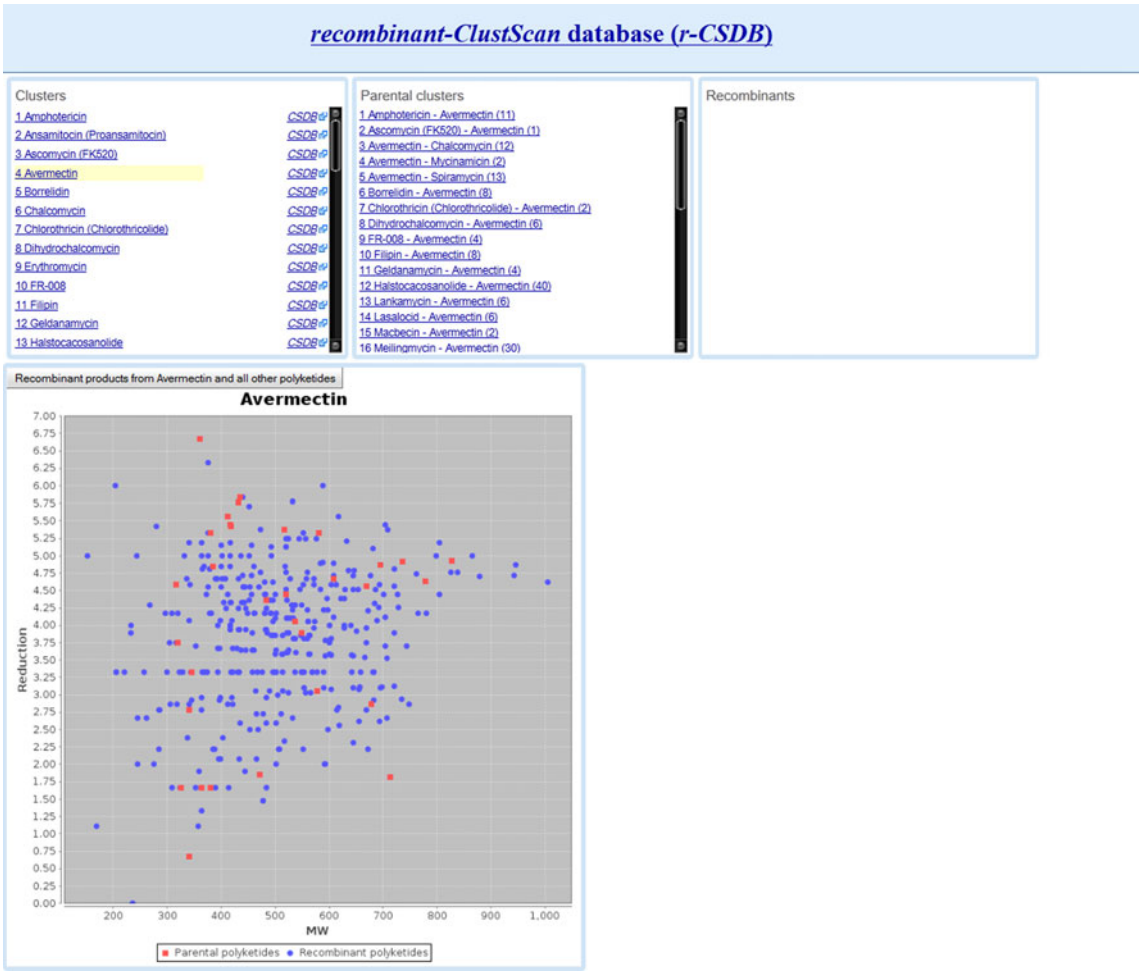

Fig. 4 Screenshot of the *r-CSDB* information about the properties of the products of all recombinants including an avermectin parent. The molecular weight and degree of reduction are shown for the recombinants (*blue*) and the parents (*red*)

engineering approaches [24]. The size of clusters (50-150 kb of DNA) is within the reach of DNA synthesis approaches and, if many recombinants were required, it would be possible to construct them economically by using common DNA segments for different recombinants [10, 18].

Acknowledgments This work was supported by the grant 09/5 (to D.H.) 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 D.H. and J.C.). It was also supported by the Leverhulme Trust; Japanese Bio-Industry Association and The School of Pharmacy, University College London (to P.F.L.).

References

1. Anand S, Prasad MV, Yadav G, Kumar N, Shehara J, Ansari MZ, Mohanty D (2010) SBSPKS: structure-based sequence analysis of polyketide synthases. Nucleic Acids Res 38:W487–W496. doi:10.1093/nar/gkq340
2. Ansari MZ, Yadav G, Gokhale RS, Mohanty D (2004) NRPS-PKS: a knowledge-based resource for analysis of NRPS/PKS megasynthases. Nucleic Acids Res 32:405–413. doi:10.1093/nar/gkh359
3. Blažič M, Lisfi M, Starcevic A, Baranasic D, Goranovič D, Fujs Š, Kuščer E, Kosec G, Petković H, Cullum J, Hranueli D, Zucko J (2012) Annotation of modular PKS and NRPS gene clusters in the genomic DNA of *Streptomyces tsukubaensis* NRRL18488. Appl Environ Microbiol 78:8183–8190. doi:10.1128/AEM.01891-12
4. Boddy CK, Christopher N (2013) ClusterMine360: a database of microbial PKS/NRPS biosynthesis. Nucleic Acid Res 41(Database issue):D402–D407. doi:10.1093/nar/gks993
5. Caboche S, Pupin M, Leclère V, Fontaine A, Jacques P, Kucherov G (2008) NORINE: a database of nonribosomal peptides. Nucleic Acids Res 36:D326–D331. doi:10.1093/nar/gkm792
6. Demain AL, Sanchez S (2009) Microbial drug discovery: 80 years of progress. J Antibiot (Tokyo) 62:5–16. doi:10.1038/ja.2008.16
7. Donadio S, Monciardini P, Sosio M (2007) Polyketide synthases and nonribosomal peptide synthetases: the emerging view from bacterial genomics. Nat Prod Rep 24:1073–1109. doi:10.1039/B514050C

8. Hertweck C (2009) The biosynthetic logic of polyketide diversity. Angew Chem Int Ed Engl 48:4688–4716. doi:10.1002/anie.200806121
9. Hranueli D, Cullum J, Basrak B, Goldstein P, Long PF (2005) Plasticity of the *Streptomyces* genome—evolution and engineering of new antibiotics. Curr Med Chem 12:1697–1704. doi:10.2174/0929867054367176
10. Hranueli D, Starcevic A, Zucko J, Rojas JD, Diminic J, Baranasic D, Gacesa R, Padilla G, Long PF, Cullum J (2013) Synthetic biology: a novel approach for the construction of industrial microorganisms. Food Technol Biotechnol 51:3–11
11. Ichikawa N, Sasagawa M, Yamamoto M, Komaki H, Yoshida Y, Yamazaki S, Fujita N (2013) DoBISCUIT: a database of secondary metabolite biosynthetic gene clusters. Nucleic Acids Res 41(Database issue):D408–D414. doi:10.1093/nar/gks1177
12. Jenke-Kodama H, Dittmann E (2009) Bioinformatic perspectives on NRPS/PKS megasynthases: advances and challenges. Nat Prod Rep 26:874–883. doi:10.1039/B810283J
13. Johnston C, Ibrahim A, Magarvey N (2012) Informatic strategies for the discovery of polyketides and nonribosomal peptides. Med Chem Commun 3:932–937. doi:10.1039/c2md20120h
14. Kalyaanamoorthy S, Chen YP (2011) Structure-based drug design to augment hit discovery. Drug Discov Today 16:831–839. doi:10.1016/j.drudis.2011.07.006
15. Kopp F, Marahiel MA (2007) Where chemistry meets biology: the chemoenzymatic synthesis of nonribosomal peptides and polyketides. Curr Opin Biotechnol 18:513–520. doi:10.1016/j.copbio.2007.09.009
16. Li MHT, Ung PMU, Zajkowski J, Garneau-Tsodikova S, Sherman DH (2009) Automated genome mining for natural products. BMC Bioinform 10:185. doi:10.1186/1471-2105-10-185
17. Medema MH, Blin K, Cimermancic P, de Jager V, Zakrzewski P, Fischbach MA, Weber T, Takano E, Breitling R (2011) antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. Nucleic Acids Res 39:W339–W346. doi:10.1093/nar/gkr466
18. Neumann H, Neumann-Staubitz P (2010) Synthetic biology approaches in drug discovery and pharmaceutical biotechnology. Appl Microbiol Biotechnol 87:75–86. doi:10.1007/s00253-010-2578-3
19. Rausch C, Weber T, Kohlbacher O, Wohlleben W, Huson DH (2005) Specificity prediction of adenylation domains in nonribosomal peptide synthetases (NRPS) using transductive support vector machines (TSVMs). Nucleic Acids Res 33:5799–5808. doi:10.1093/nar/gki885
20. 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 39:W362–W367. doi:10.1093/nar/gkr323
21. Sattely ES, Fischbach MA, Walsh CT (2008) Total biosynthesis: in vitro reconstitution of polyketide and nonribosomal peptide pathways. Nat Prod Rep 25:757–793. doi:10.1039/b801747f
22. 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. doi:10.1093/nar/gkn685
23. Starcevic A, Diminic J, Zucko J, Elbekali M, Schlosser Z, Lisfi M, Vukelic A, Long PF, Hranueli D, Cullum J (2011) A novel docking domain interface model that can predict recombination between homoeologous modular biosynthetic gene clusters. J Ind Microbiol Biotechnol 38:1295–1304. doi:10.1007/s10295-010-0909-0
24. Starcevic A, Wolf K, Diminic J, Zucko J, Trninic Ruzic I, Long PF, Hranueli D, Cullum J (2012) Recombinatorial biosynthesis of polyketides. J Ind Microbiol Biotechnol 39:503–511. doi:10.1007/s10295-011-1049-x
25. Tae H, Kong EB, Park K (2007) ASMPKS: an analysis system for modular polyketide synthases. BMC Bioinform 8:327. doi:10.1186/1471-2105-8-327
26. Weber T, Rausch C, Lopez P, Hoof I, Gaykova V, Huson DH, Wohlleben W (2009) CLUSEAN: a computer-based framework for the automated analysis of bacterial secondary metabolite biosynthetic gene clusters. J Bacterol 140:13–17. doi:10.1016/j.jbiotec.2009.01.007
27. Weininger D (1988) SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J Chem Inf Comput Sci 28:31–36. doi:10.1021/ci00057a005
28. Wong FT, Khosla C (2012) Combinatorial biosynthesis of polyketides-a perspective. Curr Opin Chem Biol 16:117–123. doi:10.1016/j.cbpa.2012.01.018
29. Yadav G, Gokhale RS, Mohanty D (2003) SEARCHPKS: a program for detection and analysis of polyketide synthase domains. Nucleic Acids Res 31:3654–3658. doi:10.1093/nar/gkg607
30. Zazopoulos E, Huang K, Staffa A, Liu W, Bachmann BO, Nonaka K, Ahlert J, Thorson JS, Shen B, Farnet CM (2003) A genomics-guided approach for discovering and expressing cryptic metabolic pathways. Nat Biotechnol 21:187–190. doi:10.1038/nbt784
31. Zotchev SB, Stepanchikova AV, Sergeyko AP, Sobolev BN, Filimonov DA, Poroikov VV (2006) Rational design of macrolides by virtual screening of combinatorial libraries generated through in silico manipulation of polyketide synthases. J Med Chem 49:2077–2087. doi:10.1021/jm051035i
32. Zucko J, Starcevic A, Diminic J, Elbekali M, Lisfi M, Long PF, Cullum J, Hranueli D (2010) From DNA sequences to chemical structures—methods for mining microbial genomic and metagenomic datasets for new natural products. Food Technol Biotechnol 48:234–242
33. Zucko J, Cullum J, Hranueli D, Long PF (2011) Evolutionary dynamics of modular polyketide synthases, with implications for protein design and engineering. J Antibiot 64:89–92. doi:10.1038/ja.2010.141
34. Zucko J, Long PF, Hranueli D, Cullum J (2012) Horizontal gene transfer drives convergent evolution of modular polyketide synthases. J Ind Microbiol Biotechnol 39:1541–1547. doi:10.1007/s10295-012-1149-2