# 7th INTERNATIONAL CONGRESS OF FOOD TECHNOLOGISTS, BIOTECHNOLOGISTS AND NUTRITIONISTS

20-23 September 2011 - Opatija, Croatia

# PROCEEDINGS

*Published by:*
Croatian Society of Food Technologists, Biotechnologists and Nutritionists

*Editor –in – Chief:*
Helga Medić

*Editorial board:*
Rajka Božanić, Danijela Bursać Kovačević, Draženka Komes, Tibela Landeka Dragičević, Ines Panjkota Krbavčić, Anita Slavica, Verica Dragović-Uzelac, Karin Kovačević Ganić, Božidar Šantek, Hami Alpas, Mauro Serafini, Jasminka Z. Ilich

*Printed by:*
BARIS d.o.o., Zaprešić

ISBN 978-953-99725-4-5

A CIP catalogue record for this book is available from the National and University Library in Zagreb under 779446

*Organised by:*

CROFOST - Croatian Society of Food Technologists, Biotechnologists and Nutritionists, Zagreb, Croatia

PBF - University of Zagreb, Faculty of Food Technology and Biotechnology, Zagreb, Croatia

*Co-sponsored by:*

# Proceedings

7th International Congress of
FOOD TECHNOLOGISTS, BIOTECHNOLOGISTS AND NUTRITIONISTS

# *3DIS*: AN EXPERT SYSTEM FOR *IN SILICO* DRUG DESIGN AND DISCOVERY

A. Starcevic [1], J. Diminic [1], J. Zucko [1], P. F. Long [2], J. Cullum [3], D. Hranueli [1]
[1] Faculty of Food Technology and Biotechnology, University of Zagreb, Zagreb, Croatia
[2] Pharmaceutical Science Institute, King's College London, Franklin-Wilkins Building, London, United Kingdom
[3] Department of Genetics, University of Kaiserslautern, Kaiserslautern, Germany

ABSTRACT

Bacterial natural products are important sources of chemical diversity for commercial exploitation by the pharmaceutical industry. In the past few years there has been a lot of interest in generating new compounds as potential drug candidates by manipulating the programming of biosynthetic gene-clusters *in vitro*. In order to assist this process a novel expert system for *in silico* drug design and discovery was developed (*3DIS*). The expert system has server-client architecture, with analysis being carried out on the server and a Java user interface for the client which can be a PC, Mac or Linux. *3DIS* consist of two program packages: for semi-automatic DNA sequence analysis (*ClustScan*) and for the generation of novel gene-clusters by virtual homeologous recombination (*CompGen*). *ClustScan* and *CompGen* were used to generate two specialised databases. *CSDB* is a ClustScan database of well characterised polyketide and nonribosomal peptide natural products. The database contains 170 uniformly annotated natural product gene-clusters. In contrast, *r-CSDB* is a virtual compound database for molecular modelling studies produced by the *CompGen* program package and contains more than 11,796 novel compounds. *In silico* studies are only useful if it is possible to generate strains producing them. Continuing progress in synthetic biology will improve methods to achieve that. A major issue for the pharmaceutical industry is maintaining a continuous supply of promising new leads for drug development. We propose that recombinatorial biosynthesis offers a new and exciting strategy whereby large and chemically diverse libraries of polyketides can first be screened *in silico* and then generated in the laboratory for further new lead development. Given that many polyketides are used clinically as antimicrobials, the *3DIS* comes at an important time when with ever increasing numbers of pathogens are becoming resistant to our current antibiotic armamentarium.

*K*EYWORDS: *natural products; gene-clusters; annotation; recombinatorial biosynthesis; databases*

## INTRODUCTION

Bacterial natural products are important source of chemical diversity for commercial exploitation by the pharmaceutical industry. In the last few years there has been a lot of interest in generating new compounds as potential drug candidates by manipulating the programming of biosynthetic gene-clusters *in vitro* (1).

The modular polyketide synthases (PKS), non-ribosomal peptide synthethases (NRPS) and polyketide/peptide (PKS/NRPS) hybrid gene-clusters, which can collectively be called Thiotemplate Modular Systems (TMS) (2), are gene-clusters whose protein products are involved in the biosynthesis of very important classes of compounds that have many useful biological activities (3). These modular biosynthetic gene-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 multimodular organisation. The growth of the polyketide or peptide carbon chains begins with the condensation of starter unit with the first extender unit and proceeds up to the completion of the linear polyketide, peptide or hybrid chains. Each module - minimally encoding ketosynthase (KS)/condensation (C) domains, acyltransferase (AT)/adenylation (A) domains and acyl carrier protein (ACP)/peptidyl carrier protein (PCP)

domains – is responsible for the extension of a polyketide, peptide or hybrid chain by one building block and might contain additional catalytically active domains for the modification of incorporated building blocks. The last module of a PKS, an NRPS or a hybrid gene-cluster usually ends with a thioesterase (TE) domain that is responsible for release of the product and its cyclisation to form an "aglycon". After that, post-polyketide or post-peptide enzymes complete the biosynthesis of a polyketide, peptide or hybrid final structures (for review see: 4). Modularity implies that the chemical structures of the aglycon products can be predicted from the DNA sequences by analysing the specificity and functionality of the individual domains in the encoded polypeptides (3).

A number of computer programs have been developed for the analysis of new modular PKS, NRPS or hybrid gene-clusters: *SEARCHPKS* (5), *DecipherIT™* (6), *NRPSpredictor* (7), *Biogenerator* (8), *MAPSI*, (9), ClustScan (10), *CLUSEAN* (11), *NP.searcher* (12) and *SBSPKS* (13). Some of them (*SEARCHPKS*, *MAPSI* and *SBSPKS*) have been used to structure and maintain publically available databases of polyketides, peptides and hybrid compounds. The PKSDB-NRPSDB database (14; 13) holds data on publically available polyketide and peptide gene-clusters including domain and module architecture and the chemical structures of the gene-clusters products. Another useful publically available polyketide database is the ASMPKS database (9). There is also the database of non-ribosomal peptide, the Norine database (15) that does not contain information on DNA sequences. In order to use these tools and assist this process we have developed a novel expert system, *3DIS* (an acronym of the term: "*In Silico D*rug *D*iscovery and *D*evelopment"), for the *in silico* drug design and discovery which is described here.

## MATERIALS AND METHODS

The *3DIS* has server-client architecture, with analysis being carried out on the server and a Java user interface for the client, which can be a PC, Mac or Linux machine. The CSDB (the ClusScan DataBase) and r-CSDB (the recombinant-ClustScan DataBase) data are stored in a relational database using a PostgreSQL database system. The graphical interface has been implemented using Java Server Page (JSP) technology with the Apache tomcat server (*http://tomcat.apache.org/*). They are available at *http://bioserv.pbf.hr/cms/*. The following informatics tools and languages have been used: Universal Markup Language [UML, (16)], BioSQL v1.29 (*http://obda.open-bio.org/*), Perl (*http://www.perl.org/*), Java (*http://java. sun.com/*) and JavaScript (*http://javascript.internet.com/*). Chemical structures of the starter and extender building blocks were encoded as extended isomeric SMILES [*S*implified *M*olecular *I*nput *L*ine *E*ntry *S*ystem (17)] allowing display of product structures using Jmol v. 11.2.14, 2006 (*http://www. jmol.org/*) and/or ChemAxon (18).

## RESULTS AND DISCUSSION

The *3DIS* consists of two suites of programs for the semi-automatic DNA sequence analysis (*ClustScan*) (10) and for the generation of novel gene-clusters by virtual homeologous recombination (*CompGen*) (19). *ClustScan* and *CompGen* are used for the generation of two specialised databases. The *CSDB* is a ClustScan database of well characterised polyketide, peptide or hybrid natural products. On the other hand, *r-CSDB* is a virtual compound database for molecular modelling studies developed by the use of *CompGen* program package.

### *The ClustScan program package*

The ClustScan (the *Clus*ter *Scan*ner) is an integrated suite of computer programs that take a "top down" approach to the annotation of gene-clusters encoding TMSs. ClustScan can analyse the large amounts of data produced by sequencing projects (genome and metagenome data sets) and can generate good predictions of the most likely chemical products from these gene-clusters, allowing identification of interesting gene-clusters. Rapid progress in sequencing technology makes the use of ClustScan ever more important. Without such a tool, it is not practical to carry out sufficiently detailed analyses of the mass of data already available today. For example, even using a conservative estimate that every actinobacterial genome contains 10 TMS gene-clusters and that 1,000 sequenced genomes will soon be available, there will be 10,000 new TMS gene-clusters, potentially encoding novel chemical entities (3; 10). We have used the ClustScan program package

to analyse prokaryotic (20) and eukaryotic (21; 22) genome (3) and metagenome (10) DNA sequences.

*The CompGen program package*

The CompGen (the *Comp*ound *Gen*erator) is also an integrated programs suite written in Java. Like ClustScan, CompGen runs on a LINUX server with a Java client on the user's computer. The major goal of the CompGen suite of programs is structuring and maintaining a publically available database of the entirely novel chemical entities generated by the in silico modelling of homeologous recombination between sequenced TMS gene-clusters. Like ClustScan, CompGen also predicts the chemical structures from the in silico generated recombinants. The future role of CompGen will be to predict the biological activities of these chemical entities using computer-aided drug design technology (23). Pairwise recombination between 1,000 TMS gene-clusters should generate nearly 1,000,000 new gene-clusters, each potentially producing a novel chemical entity. Most exciting of all, when such a product looks promising in silico, a "designer bug" can be created in the laboratory to produce it (see below).

ClustScan and CompGen were used to structure and maintain publically available databases CSDB and r-CSDB containing genetic, biochemical and chemical information on the well known natural products synthesised by TMSs, as well as of predicted, entirely novel recombinant products.

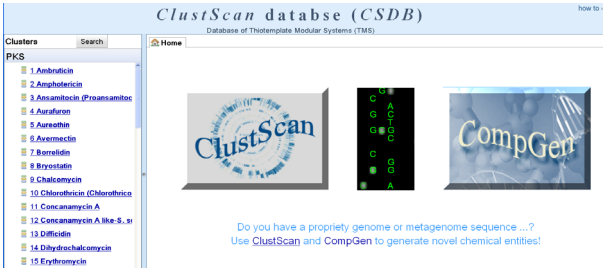
*The CSDB database*

The CSDB is a database containing genetic, biochemical and chemical information on natural products synthesised by TMSs and annotated using the ClustScan suite of programs (10). At present there are 57 PKS, 51 NRPS and 62 hybrid gene-clusters in *CSDB* (170 in total) (Fig. 1). The CSDB database contains all data starting with gene-clusters DNA sequences together with the DNA and protein sequences of annotated genes, modules and domains of TMS gene-clusters present in FASTA formats. It also contains all known polyketide and peptide building blocks in the form of isomeric SMILES, along with the programmed logic that allows total prediction of linear and partial prediction of cyclic polyketide and peptide chains and aglycons in the 2-D or 3-D forms suitable for further computer processing. Polyketide, peptide and hybrid linear chain and aglycons can be visualised using Jmol or ChemAxon. The CSDB database is fully searchable using ClustScan TMS gene-clusters annotations as well as TMS compound structures. The CSDB data can be manipulated using a number of conventional bioinformatic tools and programs.

*The r-CSDB database*

The ClustScan (10) and CompGen (Starcevic *et al*., in preparation) suites of programs were used to structure and maintain the r-CSDB database of predicted, entirely novel recombinant products that can be used for the in silico screening with the computer aided drug design technologies (23). At present there are 47 parental PKS gene-clusters, 777 cluster pairs and 20,187 recombinant gene-clusters in r-CSDB database (Fig. 2) that generates 11,796 unique compounds. Like CSDB, r-CSDB also contains all data starting with gene-clusters recombinant DNA sequence, the DNA and protein sequences of genes, modules and domains and of the recombinant gene-clusters present in FASTA formats. It also contains all known polyketide and peptide building blocks in the form of isomeric SMILES, along with the programmed logic that allows total prediction of linear and partial prediction cyclic polyketide and peptide chains and aglycons in the 2-D or 3-D forms suitable for further computer processing. Polyketide, parental and recombinant linear chain and agycons can be also visualised using Jmol or ChemAxon. The r-CSDB database is also fully searchable using CompGen (Starcevic *et al*., in preparation) suit of programs of TMS gene-cluster annotations as well as recombinant compound structures. As CSDB, the r-CSDB data can also be manipulated using a number of conventional bioinformatic tools.

A




B

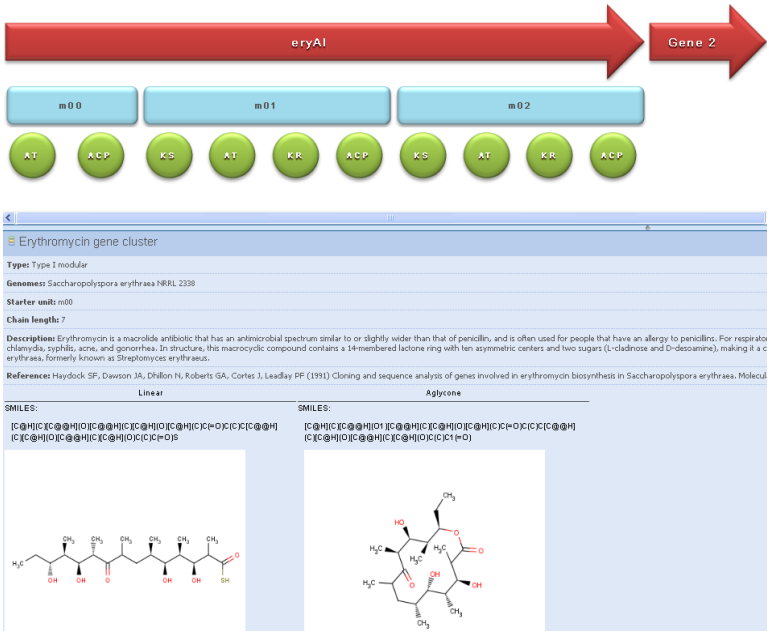
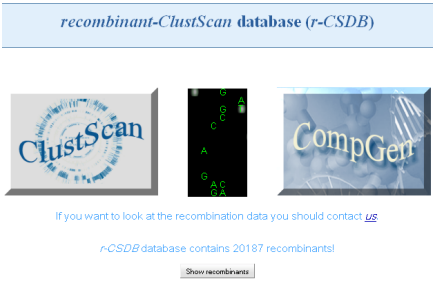



Figure 1. Screenshots of CSDB. (A) CSDB homepage and (B) the data of the erythromycin aglycon.

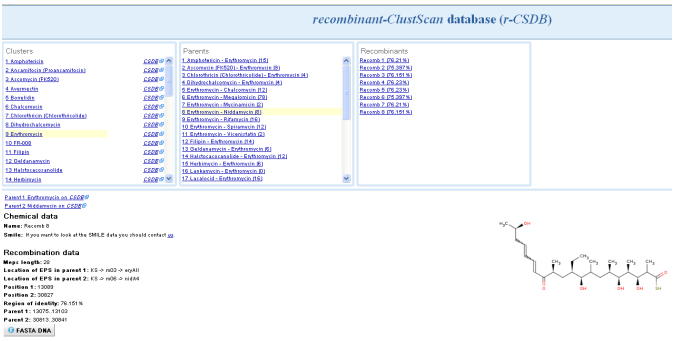
*Construction of strains producing novel chemical structures*

*In silico* studies are only useful if it is possible to generate strains producing them. An efficient system for the selection of recombination between gene-clusters, providing an effective method to mobilize hybrid gene-clusters onto various replicons which is necessary for successful engineering *in vivo* was described (Starcevic *et al*., in preparation). The advantage of such approaches is that, when a gene-cluster is cloned, it can be used for recombination with all the existing cloned gene-clusters allowing the number of compounds produced to grow rapidly. Continuing progress in synthetic biology (24) should improve the methods of synthesizing long stretches of DNA and reduce the cost. PKS gene-clusters are of significant size (50-150 kb) so they would need comparable technology to that used for chromosome synthesis of *Mycoplasma* species (1.08 Mb).

A




B




C

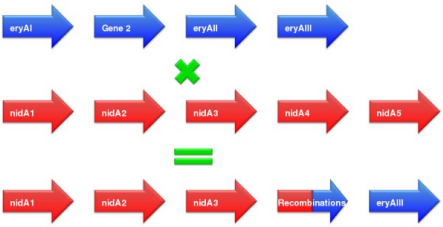



Figure 2. Screenshots of r-CSDB. (A) r-CSDB homepage, (B) the data of 7th recombinant between ery and nid gene-clusters and (C) the graphical presentation of homeologous recombination.

## CONCLUSION

A major issue for the pharmaceutical industry is maintaining a continuous supply of promising new leads for drug development. We propose that recombinatorial biosynthesis offers a new and exciting strategy whereby large and chemically diverse libraries of polyketides can first be screened *in silico* and then generated in the laboratory for further new lead development. Given that many polyketides are used clinically as antimicrobials, the *3DIS* comes at an important time, with ever increasing numbers of pathogens becoming resistant to our current antibiotic armamentarium.

## ACKNOWLEDGEMENTS

This work was supported by grant (058-0000000-3475 to D.H.) from the Ministry of Science, Education and

Sports, 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.).

*REFERENCES*

1. M.H. Medema, R. Breitling, R. Bovenberg, E. Takano, Nat. Rev. Microbiol., 9 (2011) 131-137.
2. S. Donadio, P. Monciardini, M. Sosio, Nat. Prod. Rep., 24 (2007) 1073-1109.
3. J. Zucko, A. Starcevic, J. Diminic, M. Elbekali, M. Lisfi, P.F. Long, J. Cullum, D. Hranueli, Food Technol. Biotechnol., 48 (2010) 234-242.
4. B.O. Bachmann, J. Ravel, Methods Enzymol., 458 (2009) 181-217.
5. G. Yadav, R.S. Gokhale, D. Mohanty, Nucleic Acids Res., 31 (2003) 3654-3658.
6. E. Zazopoulos, K. Huang, A. Staffa, W. Liu, B.O. Bachmann, K. Nonaka, J. Ahlert, J.S. Thorson, B. Shen, C.M. Farnet, Nat. Biotechnol., 21 (2003) 187-190.
7. C. Rausch, T. Weber, O. Kohlbacher, W. Wohlleben, D.H. Huson, Nucleic Acids Res., 33 (2005) 5799-5808.
8. S.B. Zotchev, A.V. Stepanchikova, A.P. Sergeyko, B.N. Sobolev, D.A. Filimonov, V.V. Poroikov, J. Med. Chem., 49 (2006) 2077-2087.
9. H. Tae, E.B. Kong, K. Park, BMC Bioinformatics, 8 (2007) 327.
10. A. Starcevic, J. Zucko, J. Simunkovic, P.F. Long, J. Cullum , D. Hranueli, Nucleic Acids Res., 36 (2008a) 6882-6892.
11. T. Weber, C. Rausch, P. Lopez, I. Hoof, V. Gaykova, D.H. Huson, W. Wohlleben, J. Bacterol., 140 (2009) 13-17.
12. M.H.T. Li, P.M.U. Ung, J. Zajkowski, S. Garneau-Tsodikova, D.H. Sherman, BMC Bioinformatics, 10 (2009) 185.
13. S. Anand, M.V. Prasad, G. Yadav, N. Kumar, J. Shehara, M.Z. Ansari, D. Mohanty, Nucleic Acids Res., **38 (**2010**)** W487-496.
14. M.Z. Ansari, G. Yadav, R.S. Gokhale, D. Mohanty, Nucleic Acids Res., 32 (2004) W405-413.
15. S. Caboche, M. Pupin, V. Leclère, A. Fontaine, P. Jacques, G. Kucherov, Nucleic Acids Res., 36 (2008) D326-331.
16. J. Rumbaugh, I. Jacobson, G. Booch: *The Unified Modeling Language Reference Manual*, Addison-Wesley, Boston (2005).
17. D. Weininger, J. Chem. Inf. Comput. Sci., 28 (1988) 31-36.
18. F. Csizmadia, J. Chem. Inf. Comput. Sci., 40 (2000) 323-324.
19. A. Starcevic, J. Diminic, J. Zucko, M. Elbekali, T. Schlosser, M. Lisfi, A. Vukelic, P.F. Long, D. Hranueli, J. Cullum, J. Ind. Microbiol. Biotechnol., (2010) DOI: 10.1007/s10295-010-0909-0
20. J. Zucko, J. Cullum, D. Hranueli, P.F. Long, J. Antibiot., 64 (2011) 89-92.
21. J. Zucko, N. Skunca, T. Curk, B. Zupan, P.F. Long, J. Cullum, R. Kessin, D. Hranueli, Bioinformatics, 23 (2007) 2543–2549.
22. A. Starcevic, S. Akthar, W.C. Dunlap, J.M. Shick, D. Hranueli, J. Cullum, P.F. Long, Proc. Natl. Acad. Sci. U.S.A., 105 (2008b) 2533-2537.
23. S. Mandal, M. Moudgil, S.K. Mandal, Eur. J. Pharmacol., 625 (2009) 90-100.
24. D.G. Gibson, J.I. Glass, C. Lartigue, V.N. Noskov, R.Y. Chuang, M.A. Algire, G.A. Benders, M.G. Montague, L. Ma, M.M. Moodie, C. Merryman, S. Vashee, R. Krishnakumar, N. Assad-Garcia, C. Andrews-Pfannkoch, E.A. Denisova, L. Young, Z.Q. Qi, T.H. Segall-Shapiro, C.H. Calvey, P.P. Parmar, C.A. Hutchison 3rd, H.O. Smith, J.C. Venter, Science, 329 (2010) 52-56.