From DNA sequences to chemical structures – methods for mining microbial genomic and metagenomic data sets for new natural products

Žučko, Jurica; Starčević, Antonio; Diminić, Janko; Elbekali, Mouhsine; Lisfi, Mohamed; Long, Paul F.; Cullum, John; Hranueli, Daslav

ISSN 1330-9862
(FTB-2261)

scientific note

From DNA Sequences to Chemical Structures – Methods for Mining Microbial Genomic and Metagenomic Data Sets for New Natural Products#

*Jurica Zucko[1,3†], Antonio Starcevic[1,3†], Janko Diminic[1], Mouhsine Elbekali[3], Mohamed Lisfi[3], Paul F. Long[2], John Cullum[3] and Daslav Hranueli[1]**

[1]Faculty of Food Technology and Biotechnology, University of Zagreb, Pierottijeva 6, HR-10000 Zagreb, Croatia

[2]School of Pharmacy, University of London, 29/39 Brunswick Square, London WC1N 1AX, United Kingdom

[3]Department of Genetics, University of Kaiserslautern, Postfach 3049, DE-67653 Kaiserslautern, Germany

Received: March 19, 2009
Accepted: January 26, 2010

Summary

Rapid mining of large genomic and metagenomic data sets for modular polyketide synthases, non-ribosomal peptide synthetases and hybrid polyketide synthase/non-ribosomal peptide synthetase biosynthetic gene clusters has been achieved using the generic computer program packages *ClustScan* and *CompGen*. These program packages perform the annotation with the hierarchical structuring into polypeptides, modules and domains, as well as storage and graphical presentations of the data. This aims to achieve the most accurate predictions of the activities and specificities of catalytically active domains that can be made with present knowledge, leading to a prediction of the most likely chemical structures produced by these enzymes. The program packages also allow generation of novel clusters by homologous recombination of the annotated genes *in silico*. *ClustScan* and *CompGen* were used to construct a custom database of known compounds (*CSDB*) and of predicted entirely novel recombinant products (*r-CSDB*) that can be used for *in silico* screening with computer aided drug design technology. The use of these programs has been exemplified by analysing genomic sequences from terrestrial prokaryotes and eukaryotic microorganisms, a marine metagenomic data set and a newly discovered example of a 'shared metabolic pathway' in marine-microbial endosymbiosis.

Key words: polyketides, non-ribosomal peptides, *Actinobacteria*, homologous recombination

Introduction

It is well known that more than 50 % of drugs that are in clinical use today belong to the polyketide or non-ribosomal peptide families of natural products. As well as antibacterials, antifungals, antivirals and cytostatics, a number of immunosuppressants, antihypertensives, antidiabetics, antimalarials and chlolesterol-lowering polyketide and non-ribosomal peptide synthase-derived drugs are in clinical use. Polyketide and non-ribosomal peptide antiparasitics, coccidiostatics, animal growth promoters and natural insecticides are also used in the food

*Corresponding author; Phone: ++385 1 460 5013; Fax: ++385 1 483 6083; E-mail: dhranueli@pbf.hr

†These authors contributed equally to this work

#This work was presented at the ECI Conference 'Natural Products Discovery and Production II: Celebrating the Successes of Traditional and Novel Culture Sources' held during June 22–26, 2008 in Whistler, British Columbia, Canada

and agro-industries and should be added to this list of useful biological activities (*1–5*). Polyketides and non-ribosomal peptides are, therefore, very important classes of compounds which will continue to be sourced for novel drug discovery. The year 2009 marked the eightieth anniversary of the discovery of penicillin (*6*). In the intervening years, pharmaceutical companies and academic institutions had screened cultivable microorganisms from different habitats for their natural products. Tens of thousands of biologically active polyketides and non-ribosomal peptides with important pharmacological properties have been discovered, but only 350 antimicrobials have found their way to the market (*7*). Therefore, only 1 to 3 biologically active chemical entities per 10 000 described have found clinical use. During the last 40 years or so, the numbers of natural products reaching the market sourced from cultivable microorganisms has declined considerably. The major question now is: are there any alternative strategies to exploit the chemical potential of small-molecule natural products?

The answer might lie in the recent development of rapid and inexpensive DNA sequencing technologies, which have resulted in over 1050 sequenced bacterial genomes, with more than 3000 additional genome sequencing projects in progress [NCBI Genome, May 2010 (*8*)]. Modular polyketide synthase (PKS) and non-ribosomal peptide synthethase (NRPS) gene clusters, which can collectively be called Thiotemplate Modular Systems (TMS), have been analyzed within about 250 bacterial genomes. It has been shown that the bacteria whose genomes are smaller than 4 Mb have only a few or no TMS gene clusters. The number of gene clusters present in a genome increases rapidly with the size of the bacterial genome (*9*). Many actinomycetes, such as *Streptomyces* species, have large genomes of around 8 Mb in size. All such species sequenced to date encode from 20 to 40 gene clusters for the synthesis of secondary metabolites (*10*). For instance, the genome of *Streptomyces avermitilis* contains gene clusters encoding three well known polyketides: the antiparasitic avermectin and the antibiotics oligomycin and filipin. In addition, there is a Type III PKS (*rpp*) as well as eight PKS and eight NRPS gene clusters with unknown biosynthetic products. When the gene clusters encoded within the genomes of *Streptomyces coelicolor*, *Saccharopolyspora erythraea*, *Streptomyces griseus*, *Salinispora tropica*, *Salinispora arenicola* and *Rhodococcus jostii* are considered, there are almost 200 gene clusters for secondary metabolites within only seven genomes of which almost 150 have unknown biosynthetic products (*10*). To date hundreds of sequenced secondary metabolite gene clusters encoding unknown products have been deposited in public databases and their number is growing exponentially. It can be expected that with the development of 'next generation' sequencing technology (*11*), more than 1000 actinomycete genomes will be sequenced in the next year or two. If every sequenced genome contains about 20 secondary metabolite gene clusters of unknown biosynthetic products, more than 20 000 as yet unidentified chemical entities can be expected. The majority of these gene clusters are probably actively expressed. They have typical TMS gene, module and domain organisation with no detectable rearrangements or mutations which would result in inactive enzymes. It is reasonable to believe that there is, therefore, a significant untapped chemical diversity in these sequenced genomes and gene clusters. A major problem of traditional screening is the reisolation of known compounds (*7*). It is clear that a similar problem will eventually arise with the data mining of genome sequences. However, to date, most clusters detected in genome sequences are novel in the sense of not having a high degree of sequence similarity with known clusters. The analysis tools provided by *ClustScan* provide enough chemical information to detect close analogues of known polyketides and peptides. As DNA sequencing is much cheaper and less labour-intensive than chemical purification and characterisation (*11*), the elimination of such compounds at a DNA level makes this approach viable even if increasing data makes the occurrence of known structures common. At present, the data do not allow a reliable estimate of the total chemical diversity that can be found by a DNA sequencing approach.

Developments in metagenomics (*12*) have also opened up new opportunities to access the genomes and, hence, metabolic potential of the uncultivable majority of microorganisms. For example, in 2007 the J. Craig Venter Institute (*13*) published the results of the Sorcerer II global ocean sampling expedition to evaluate the microbial diversity in the world's oceans using molecular tools and techniques originally developed to sequence the human genome. Their data represent the largest metagenomic data set ever released into the public domain, with almost 8 million sequencing reads, comprising over 6 billion base pairs of DNA. Bioinformatic tools are now needed to rapidly mine such massive data set to hunt for gene clusters encoding entirely novel compounds that could be the source of information, which will aid the development of the medicines of the future. This is where we hope our approach will become important.

Here we describe the development of a novel approach for *in silico* drug design and discovery that links DNA sequences encoding modular enzymes that synthesize polyketide and non-ribosomal peptide natural products of clinical importance to their *in silico* medicinal chemistry. This is made possible because there is a direct correlation between the DNA sequence of the gene cluster and the chemical structure of the biosynthetic product they encode. Therefore, polyketide and non-ribosomal peptide chemical structures can be predicted with a relatively high probability from the gene cluster DNA sequences.

Materials and Methods

Bacterial gene and genome (*8*) databases were downloaded from the FTP server at NCBI onto the Bioserv server (*14*), which is the bioinformatics server of the Faculty of Food Technology and Biotechnology, University of Zagreb, Croatia and used locally as the starting point for the accession of sequenced polyketide and non-ribosomal peptide gene clusters. *BLAST* (*15*) searches of the bacterial genes and genomes were done locally using databases on the Bioserv. The bacterial polyketide and non-ribosomal peptide domain sequences were accessed from the PKSDB-NRPSDB and ASMPKS databases (*16*, *17*).

The generic computer program packages *ClustScan* and *CompGen* (licensed by Novalis, Ltd, Zagreb, Croatia) were used for gene cluster annotation and novel polyketide structure generation, respectively (*18*,*19*). Coding regions were identified using *Glimmer* (*20*) and *GeneMark* (*21*). The DNA sequences were translated using *Transeq* (*22*). Custom profile design and searching with profiles used the *HMMER* suite of programs (*23*) and the Pfam database (*24*).

A special data structure was developed to describe modular clusters. It includes a hierarchical organisation of the clusters into polypeptides, modules and domains as well as information specifying the biosynthetic order of the polypeptides and the specificity and functionality of the domains. For structuring of the custom databases *CSDB* (ClustScan *D*ata*B*ase) and *r-CSDB* (*r*ecombinant--ClustScan *D*ata*B*ase), which can be searched at the *TMSS* (*T*hiotemplate *M*odular *S*ystems *S*tudies) web portal (*25*), as well as for the development of the computer program packages *ClustScan* and *CompGen* (*18*,*19*), the following informatics tools and languages were used: Universal Markup Language (UML, *26*), BioSQL v. 1.29 (*27*), Perl (*28*), Java (*29*) and JavaScript (*30*). Chemical structures of the starter and extender building blocks were designed as extended isomeric SMILES (*S*implified *M*olecular *I*nput *L*ine *E*ntry *S*ystem, *31*) using Jmol (*32*) and/or ChemAxon (*33*).

Results and Discussion

Two generic computer program packages named *ClustScan* and *CompGen* (*18*,*19*) were developed. *ClustScan* and *CompGen* are acronyms for '*Clust*er *Scan*ner and *Comp*ound *Gen*erator' and are written in Perl (*28*), Java (*29*) and JavaScript (*30*). *ClustScan* and *CompGen* run on a *Linux* server with a Java client on the user's computer compatible with *Windows*, *MacIntosh* and *Linux* operating systems. The *ClustScan* program package lets users perform rapid data mining (*19*). It allows convenient, rapid, semi-automatic annotation of modular biosynthetic clusters with knowledge-based predictions of enzyme properties: domains, activities and domain specificities. The main features of the *ClustScan* program package are shown in Fig. 1. Namely, DNA sequences can be imported from any file supported by *ReadSeq* (*34*); either from the web or the user's hard disk. Once imported, the DNA sequence is automatically translated into amino acid sequence using *Transeq* (*22*). Genes can be localized within the translated protein sequences by *Glimmer* (*20*) or *GeneMark* (*21*). The domains of polyketide or non-ribosomal peptide gene clusters can then be identified within these genes using the *HMMER* suite of programs (*23*) with stringent or relaxed parameters, exploiting protein profiles from the Pfam database (*24*) or customized profiles. The final annotation can be exported as *EMBL*, *GenBank* or *XML* file for use in other applications (*18*,*19*), or for the upload to *CSDB* or *r-CSDB* databases (*25*).

The accuracy of prediction domain activity and specificity, and the speed and convenience of annotating large DNA data sets make *ClustScan* a useful tool. The accuracy of prediction was tested on a number of well--known gene clusters; *e.g.* the erythromycin gene cluster (GenBank accession number AY771999) (Fig. 2). The 'Cluster' editor (Fig. 2a) and the 'Workspace' window (Fig. 2b) show all three erythromycin genes together with all their domains. The annotation of the cluster is carried out using the 'Cluster' editor, which shows genes of the cluster in a simple cartoon form, hence the term semi-automatic. When a gene is selected, the domains are shown. The 'Workspace' window shows the results of the annotation in a tree format in which branches can be opened up or collapsed. This is useful to obtain an overview. However, the overall annotation of the erythromycin biosynthetic genes can be best viewed in the 'Annotation' editor (Fig. 2c) where domains are assigned to modules and modules to genes. For example, *ClustScan* correctly predicted that inactive KR domain in module 3 is responsible for S β-carbon stereochemistry. From the annotated data, *ClustScan* can automatically predict a structure of the predicted linear 6-deoxyerythronolide (6dEB) polyketide-derived chain using building blocks saved as isomeric SMILES (*31*) in a two- or three-dimensional form (2-D or 3-D) suitable for further computer processing. Based on the prediction of a linear polyketide backbone, the program also automatically produces a 2-D or 3-D cyclic structure of the erythromycin aglycon (6-dEB). The *ClustScan* programme package is used to upload the public database *CSDB* developed as a web application using AJAX technology (Google Web Toolkit, *35*), which can be searched at *TMSS* web portal (*25*).

The J. Craig Venter Institute metagenomic data set (*13*) was also analyzed. This revealed a potential hybrid PKS/NRPS gene cluster of about 50 kb in size. The cluster begins by encoding a peptide loading domain, followed by three putative PKS modules and seven putative NRPS modules, ending with an NRPS thioesterase domain. When the location of the domains with respect to the predicted polypeptides is examined, the cluster appears to have been inactivated by several frameshifts. However, because of the technical problems with metagenomic DNA samples, the quality of the sequence is usually much lower than for genome projects. This makes it likely that the cluster is indeed functional, and the graphical overview provided by *ClustScan* makes correction for sequencing errors relatively easy. Based on this analysis, it is suggested that there are three apparent frameshifts and one false stop codon due to sequencing errors. Thus, it seems likely that there are three genes rather than the seven indicated by both *GenMark* and *Glimmer* analysis. In the case of two of the putative polyketide modules, no acyl transferase domains are recognized, but there are unassigned regions in the putative protein of appropriate sizes and locations for acyl transferase domains. Thus, the program allows rapid scanning of metagenomic data sets and makes it easy to identify potential sequencing errors and interesting features of gene clusters. With the growing importance of metagenomic data for drug discovery programs, *ClustScan* helps to eliminate major bottlenecks in the analysis, showing the speed and convenience of annotating large DNA data sets (*18*,*19*). The current version of *ClustScan* can be downloaded from the Bioserv (*14*) web site. Most of the current metagenomic DNA sequences are in short contigs (<10 kb), which makes detection of complete modular biosynthetic clusters impossible. However, rapid technical progress should result in future data sets

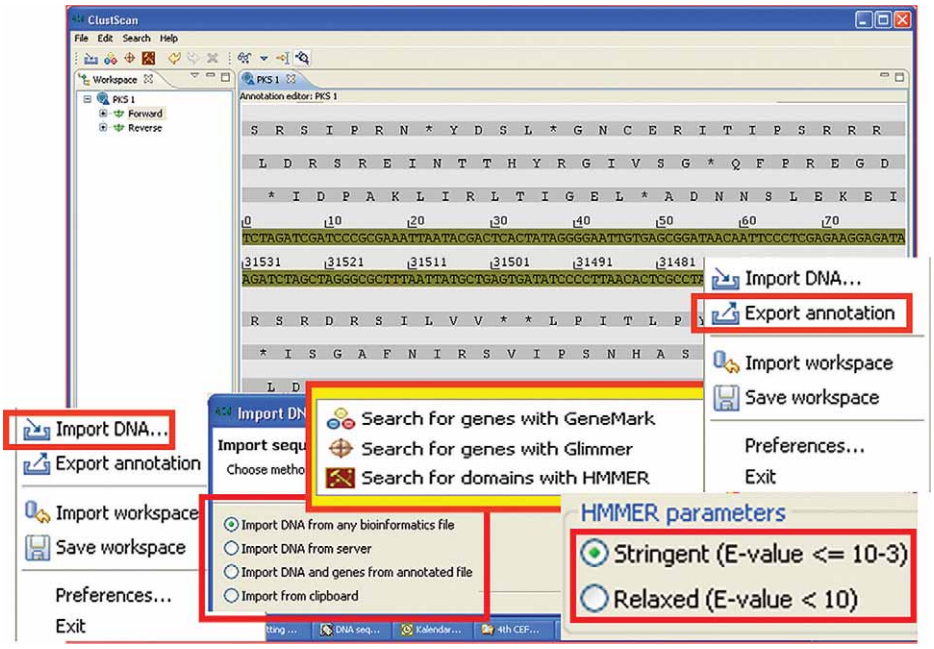

Fig. 1. The main features of the *ClustScan* program package. The annotation editor window at the back shows the DNA sequence with *in silico* translation in all six reading frames. The file menu allows import and export of data and the option to import DNA is boxed. Clicking on this item gives a second menu which gives options for import of DNA sequence from different sources. The general analysis tools to find domains (*HMMER*) or genes (*GeneMark* or *Glimmer*) are given in a further menu. When *HMMER* searches are undertaken, the menu offers a choice of stringent or relaxed parameters. Once annotation has been successfully completed, an annotated file in standard format can be exported using the option in the file menu (boxed)

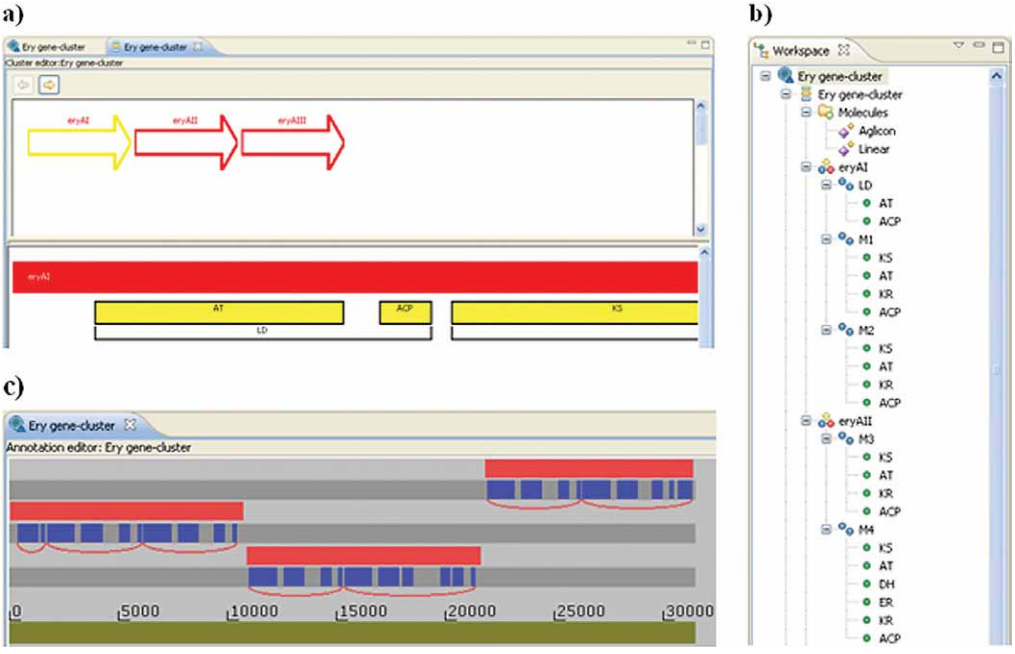

Fig. 2. The 'Cluster' editor (a), the 'Workspace' window (b) and the 'Annotation' editor (c) are shown using the example of the erythromycin gene cluster (modified from reference *21* with the permission of publisher)

with much larger contig sizes, which will reveal novel clusters.

In 2007, Peter Leadlay's group at the University of Cambridge published the DNA sequence of the *Saccharopolyspora erythraea* genome (*36*), the producer of erythromycin. The *Streptomyces scabies* genome was also sequenced and is publicly available, although not yet published (*37*). These genome sequences were used to evaluate *ClustScan* regarding the speed and convenience of annotating large DNA data sets (Fig. 3). TMS gene clusters in both genomes were annotated in about three hours by a person familiar with the *ClustScan* program package. In principle, it would be possible to achieve the same results by using individual tools such as *BLAST*

(*15*), *HMMER* (*23*) and *GenMark* (*21*). The results could then be combined with an extensive knowledge of the literature about the activity and specificity of domains to produce predictions of module specificity. These could be combined with chemical expertise to produce a predicted chemical structure of the product. This would take at least three to four weeks, but probably much longer. It would also require a much greater degree of bioinformatic expertise and a much more detailed knowledge of the literature than needed for using *ClustScan*. A further serious problem that should not be underestimated is that human analysis of such complex data will frequently lead to errors that are difficult to detect as the details of each step of the analysis will usually not be documented. Therefore, use of the *ClustScan* program package should significantly increase the productivity of scientists when faced with the annotation of complex biosynthetic processes from large genome and metagenome data sets (*18*,*19*).

ClustScan is mainly designed for use with bacterial sequences (*18*,*19*). However, the generic nature of the *ClustScan* program package has also been tested for the analysis of sequences from lower eukaryotes, where intron prediction is often difficult. An example is provided by the slime mould *Dictyostelium discoideum*, which was found to encode 45 PKS genes, the largest number yet described in a single organism (*38*). It was possible to use local *HMMER* profiles for the predicted PKS domains to effectively recognize segments of the domains split by introns. When such an analysis is carried out, a PKS gene displays a characteristic signature with parts of the protein domains in the correct order, but with gaps due to introns in between. The view in the 'Annotation' editor window allows easy recognition of genes and the coordinates of the domain segments help in detecting the intron boundaries (Fig. 4). Although *ClustScan* is mainly designed for the annotation of gene clusters encoding modular biosynthetic enzymes (*18*,*19*), it can also be used for annotating other genes by loading appropriate *HMMER* protein profiles. For instance, we have used seven profiles to find and annotate shikimic acid pathway genes in a marine animal *Nematostella vectensis*, challenging the traditional view that animals are considered to lack this pathway as evidenced by their dietary requirement for shikimate-derived aromatic amino acids. Using *ClustScan*, these genes either had codon usage that strongly implicates horizontal gene transfer from a bacterial donor, or contained introns suggesting a dinoflagellate ancestor (*39*).

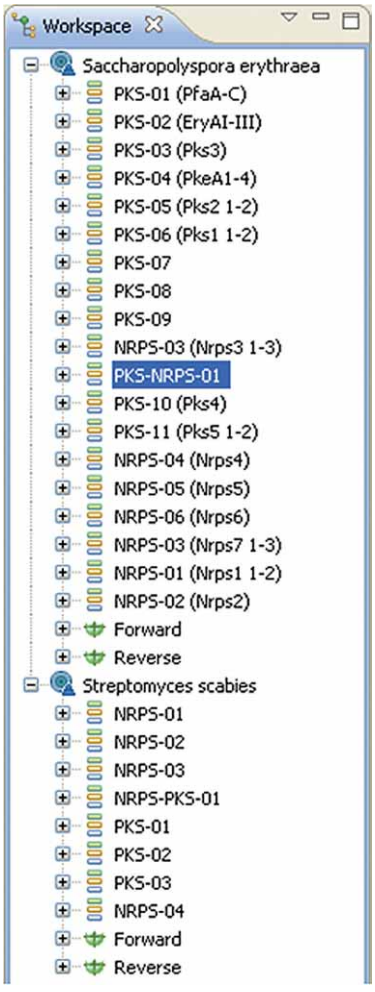

Fig. 3. The annotation of the TMS gene clusters in *Saccharopolyspora erythraea* and *Streptomyces scabies* genomes

In contrast, the major objective of *CompGen* (the details of which will be published elsewhere, but have been discussed in reference *18*) is structuring and maintenance (*i.e.* updating as improved knowledge becomes available) of the propriety custom database, *r-CSDB* (*25*), of the entirely novel chemical entities generated by *in silico* modelling of the homologous recombination between sequenced gene clusters. Prediction of the chemical product of recombinants is more difficult for NRPS clusters than for PKS clusters, but is also possible in many cases. The *r-CSDB* was also developed as a web application using AJAX technology (*35*) and can also be searched at *TMSS* web portal (*25*). Like *ClustScan*, *CompGen* also predicts the most likely chemical structures of the *in silico* generated recombinants. The future of *CompGen* will be to produce structures that can be used for *in silico* prediction of biological activity, for instance, modelling the interaction with protein targets using computer aided drug design technology (*40*). In nature, homologous recombination requires about 30 bp of identical sequence called MEPS (*M*inimal *E*fficient *P*airing *S*equence), flanked by around 200 bp of sequence sharing greater than 75 % similarity (*41*,*42*). The *CompGen* recombination program finds identity and similarity in two sequences and links their coordinates to a custom database so that domains, linkers and dockers involved in the recombination process can be recognized. Why is *Streptomyces* a good choice for generating recombinants? *Streptomyces* DNA has a high G+C content and this introduces constraints on codon usage. This results in longer stretches of sequence identity and similarity than for DNA sequence pairs with a less extreme G+C content, which encode homologous modules of PKS and NRPS clusters. *Streptomyces* have no *mut*SL genes which prevent recombination between sequences with mismatches (*43*). In addition, *Streptomyces* chromosomes and plasmids are often linear; single crossovers would, thus, result in two linear recombinant molecules rather than in a single fusion molecule that would occur with circular parent molecules

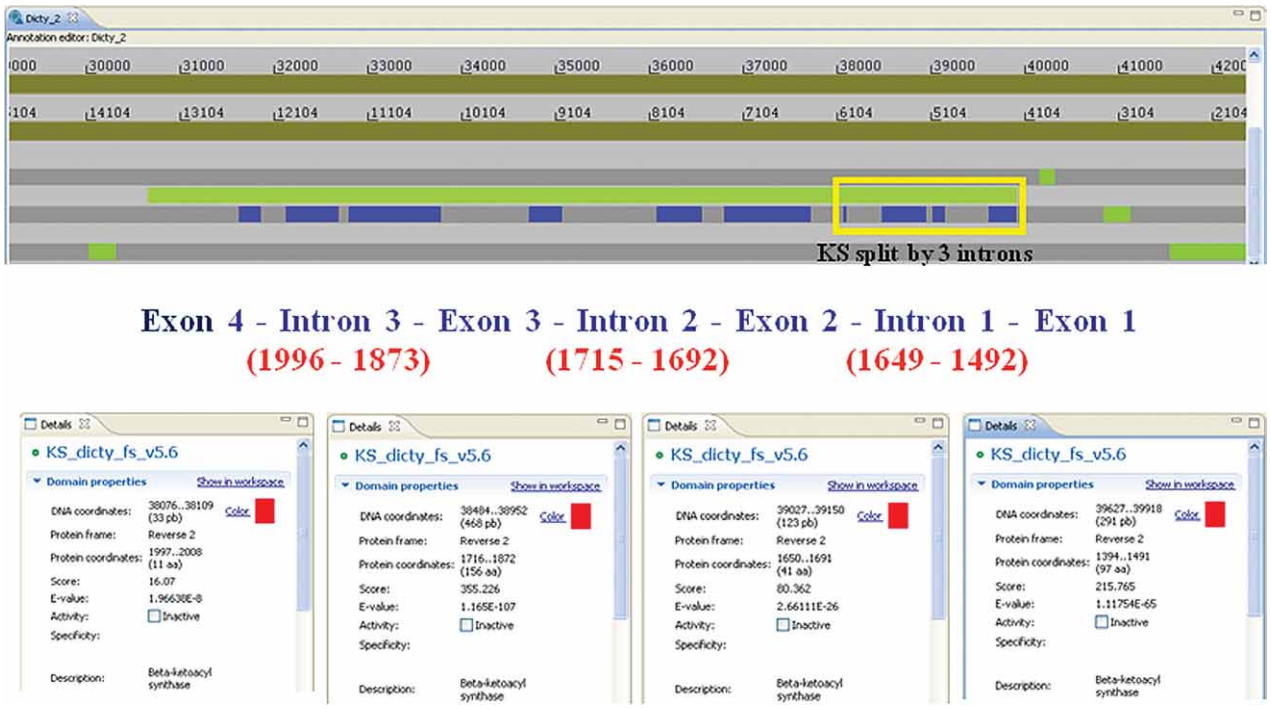

Fig. 4. The *Dictyostelium discoideum* gene *pks*29. The annotation editor window has been zoomed out to give an overview of the gene so that the DNA and amino acid sequences are not shown; as all the gene is encoded by the lower strand, the viewing option has been set for only the three reading frames on this strand. The *HMMER* hits to PKS domains are shown in the dark bars within the reading frame. There are four hits to the ketosynthase (KS) domain (boxed). By right clicking on these hits, it is possible to obtain the detail windows, which show coordinates of hits as well as the region of the profile, which is aligned with the hits. These data show that the four hits correspond to successive regions of the profile without any gaps, which indicates that the KS domain is split by three introns. The exon and intron coordinates can be deduced from the coordinates of the KS domains recognised by the local profile

(*44*,*45*). Circular molecules must undergo a double cross over to generate two recombinant molecules and this would be expected to be a much rarer event. However, the recombination process in *Streptomyces* is not well understood. *Streptomyces* do have *rec*A genes for homologous pairing, but seemingly do not have homologues of *rec*BCD or *add*AB genes (*46*,*47*). However, there are 4 conserved potential DNA helicases in the genomes of *S. coelicolor* A3(2), *S. avermitilis* and *Sacc. erythraea* (*10*). It is intriguing to speculate that one of these may encode an enzyme with similar activities. Nevertheless, by using parameters of 27 bp MEPS identity and over 75 % similarity in the neighbouring 200 bp, 275 putative recombination sites between 16 well described clusters were obtained, which gives an average of 2–3 recombination sites per cluster pair. Therefore, if 1000 gene clusters were available, recombination between these cluster pairs would generate 2–3 million recombinants. Preliminary data indicate that 30–50 % of recombinants have a PKS structure compatible with encoding a functional product so that it should be theoretically possible to generate approximately a million novel chemical entities.

As an example, two PKS gene clusters can be taken: one encoding the 16-membered macrolide niddamycin and the other encoding the 14-membered macrolide erythromycin (GenBank accession numbers AF016585 and AY771999). The niddamycin gene cluster contains five genes encoding seven modules that incorporate eight building blocks into the niddamycin aglycone. In contrast, the erythromycin gene cluster contains three genes encoding six modules that incorporate seven building blocks into the erythromycin aglycone. The computer program for modelling *in silico* homologous recombination between these two gene clusters predicts 4 putative recombination sites, *i.e.* eight novel recombinant gene clusters. In one of them, the recombination occurs between the *nid*AIV and *ery*AII genes, which contain a common EPS (*E*fficient *P*airing *S*equence) of 28 nucleotides within 200 bp of over 75 % sequence similarity. The recombination program recognizes the start and the stop coordinates of the predicted EPS and localizes these within the KS6 domain of *nid*AIV and the KS3 domain of *ery*AII. Recombination would generate two novel gene clusters, Rec-1 and Rec-2. The *CompGen* recombination program can, from the activities and specificities of polyketide domains and from the building blocks saved as isomeric SMILES (*31*), predict the putative 12- and 18--membered linear chains and aglycones that could potentially be produced by the resulting hybrid synthases.

A tool that helps the analysis of new gene clusters is the PKSDB-NRPSDB database (*16*) (*SEARCHPKS*, *48*), which holds data on publically available polyketide and peptide gene clusters including domain and module architecture and the chemical structures of the gene cluster products. It allows users to input protein sequences to be used in *BLAST* searches (*15*) to identify domains and to find the closest matching sequences in the database. This allows prediction, for example, of whether an

AT domain uses malonyl-CoA or methylmalonyl-CoA as a substrate (*i.e.* whether a C2 or C3 unit is incorporated into the polyketide). Another useful resource is the ASMPKS database (*17*) with a tool *MAPSI*, which uses a similar analysis methodology to *SEARCHPKS*, but integrates it with a graphical display that shows the presence of domains in genes so that modules can be easily recognized. It also allows display of a predicted linear polyketide chain product for which the user has to select starter and extender units from a list. However, the stereochemistry of the product is not considered.

The company ECOPIA has also developed a software tool, *DecipherIT*™ (*49*), which helps annotation of new gene clusters based on the comparison with a database of known clusters. A slightly different tool is the *Biogenerator* program (*50*) that allows construction of new polyketides based on the known modules *in silico*. The programs described above all depend on comparisons with sequences in a database to analyze new gene clusters. They are very suitable for analysing clusters closely related to well characterized gene clusters. However, all the programs mentioned above use predictions based on similarity searches and do not predict domain activities/specificities and stereochemistry. A similarity approach using *BLAST* (*15*) suffers from the problems that the weighing of each residue is independent of the position and that insertion or deletion of residues is not handled well. In contrast, the hidden Markov model (HMM) approach used in the *HMMER* suite of programs (*23*) deals with each amino acid position separately and, thus, distinguishes between critical highly conserved residues and less important positions. In addition, the HMMs consider insertion and deletion states and assign them probability. The *hmmalign* program of the *HMMER* suite aligns sequences with respect to the HMM profile, allowing accurate identification of particular residue positions. Such HMM approaches were used to predict substrate specificity of non-ribosomal peptide adenylation domains, which has been incorporated into a program package called the *NRPSpredictor* (*51*).

We compared the performance of *ClustScan* (*19*) to that of the *SEARCHPKS* and *MAPSI* prediction programs implemented within PKSDB-NRPSDB and ASMPKS databases (*16,17*; Table 1). The *SEARCHPKS* program (*48*) requires protein sequence input so that preparatory work to find genes and translate the DNA is necessary. The *MAPSI* program (*17*) works well with genome length DNA sequences, but the gene finder is not effective for high G+C DNA of typical cluster length (approx. 100 kb). The *SEARCHPKS* and *MAPSI* programs are reasonably effective in recognising AT-specificity; this probably reflects the fact that AT specificity correlates well with phylogenetic trees (*52*) so that, in addition to critical functional amino acids, there are other amino acids that differ for evolutionary reasons. However, they do not attempt to predict the hydroxyl or methyl group stereochemistries that are determined by the KR domain. They also do not predict inactive reduction domains. It is likely that the use of *BLAST* (*15*) searches will not be capable of recognising such subtle properties of domains. In *ClustScan* the use of *HMMER* (*23*) profiles and amino acid fingerprints has overcome these problems. Any analysis program will encounter problems with 'unusual' clusters. The ability to edit automatic predictions and override them is therefore important. The *SEARCHPKS* and *MAPSI* programs do not allow this. A further restriction of these two programs is that they cannot output the results of analyses in a standard form such as an annotation file (GenBank or EMBL format) or the results of chemical structure predictions in a format such as SMILES (*31*).

Conclusions

ClustScan is an integrated suite of computer programs that takes a 'top down' approach to annotation. It makes it possible to analyse the large amounts of data produced by sequencing projects (genome and metagenome data sets) and produce good predictions of the chemical products allowing identification of interesting clusters. Without such a tool, it is not practicable to carry out sufficiently detailed analyses of the mass of data already available today. Rapid progress in sequencing technology makes the use of *ClustScan* and *CSDB* even more important. For example, even using the conservative assumptions that an actinomycete with a large genome (*e.g. Streptomyces* species) contains on average 20 TMS gene clusters with unknown biosynthetic products and that 1000 sequenced genomes will be available in the near future, the identification of 20 000 new gene

Table 1. The comparison of the *ClustScan*, *SEARCHPKS* and *MAPSI* systems

Features	*ClustScan*	*SEARCHPKS*	*MAPSI*
DNA input	Yes	Only protein	Yes[a]
Recognises AT specificity	Yes	Yes	Yes
Recognises KR stereochemistry	Yes	No	No
Recognises inactive KR domains	Yes	No	No
Recognises inactive DH domains	Yes	No	No
Recognises inactive ER domains	Yes	No	No
Editing of predictions	Yes	No	No
Exports annotation file	Yes	No	No
Predicts chemical structure	Yes	No	Yes[b]
Exports chemical structure in standard format	Yes	No	No

[a]needs long DNA sequences (>200 kb for high G+C content) for accurate gene identification

[b]limited structural prediction without editing facilities

cluster DNA sequences can be expected. At present, most of the clusters revealed by genome projects (*e.g. S. avermitilis*) are novel, so it is likely that most of these new clusters will correspond to novel chemical entities. The *CompGen* program suite has also been developed. It will allow scientists to search our propriety custom database *rCSDB* of entirely novel chemical structures generated by *in silico* homologous recombination and to test these structures for potential biological activity using computer-aided drug design technology. Pairwise recombination between 1000 TMS gene clusters could generate 1 000 000 new recombinant gene clusters potentially producing novel chemical entities. Most important of all, when such a product looks promising *in silico*, a 'designer bug' can be created in the laboratory to produce it.

Addendum in Proof

Since sending to press, two additional tools that help the analysis of new gene clusters *CLUSEAN* (*53*) and *NP.searcher* (*54*) have been published but not compared with the performance of *ClustScan* (*19*).

Acknowledgements

This work was supported by grant 058-0000000-3475 (to D.H.) from the Ministry of Science, Education and Sports, the Republic of Croatia and by a cooperation grant of the German Academic Exchange Service (DAAD) and the Ministry of Science, Education and Sports, the Republic of Croatia (to D.H. and J.C.). Additional support was received from the Leverhulme Trust, Japanese Bio-Industry Association and the School of Pharmacy, University of London (to P.F.L.).

References

1. R. Finking, M.A. Marahiel, Biosynthesis of non-ribosomal peptides, *Ann. Rev. Microbiol. 58* (2004) 453–488.
2. D. Hranueli, J. Cullum, B. Basrak, P. Goldstein, P.F. Long, Plasticity of the *Streptomyces* genome – Evolution and engineering of new antibiotics, *Curr. Med. Chem. 12* (2005) 1697– 1704.
3. E.S. Sattely, M.A. Fischbach, C.T. Walsh, Total biosynthesis: *In vitro* reconstitution of polyketide and nonribosomal peptide pathways, *Nat. Prod. Rep. 25* (2008) 757–793.
4. K.J. Weissman, P.F. Leadlay, Combinatorial biosynthesis of reduced polyketides, *Nat. Rev. Microbiol. 3* (2005) 925–936.
5. H. Jenke-Kodama, E. Dittmann, Bioinformatic perspectives on NRPS/PKS megasynthases: Advances and challenges, *Nat. Prod. Rep. 26* (2009) 874–883.
6. A. Fleming, On the antibacterial action of cultures of a *Penicillium*, with special reference to their use in the isolation of *B. influenzae*, *Brit. J. Exp. Pathol. 10* (1929) 226–236.
7. A.L. Demain, Antibiotics: Natural products essential to human health, *Med. Res. Rev. 29* (2009) 821–842.
8. Prokaryotic Projects, National Center for Biotechnology Information, Bethesda, MD, USA (*http://www.ncbi.nlm.nih.gov/*).
9. S. Donadio, P. Monciardini, M. Sosio, Polyketide synthases and nonribosomal peptide synthetases: The emerging view from bacterial genomics, *Nat. Prod. Rep. 24* (2007) 1073–1109.
10. M. Nett, H. Ikeda, B.S. Moore, Genomic basis for natural product biosynthetic diversity in the actinomycetes, *Nat. Prod. Rep. 26* (2009) 1362–1384.
11. S.C. Schuster, Next-generation sequencing transforms today's biology, *Nat. Methods, 5* (2008) 16–18.
12. W.C. Dunlap, M. Jaspars, D. Hranueli, C.N. Battershill, N. Perić-Concha, J. Zucko, S.H. Wright, P.F. Long, New methods for medicinal chemistry – Universal gene cloning and expression systems for production of marine bioactive metabolites, *Curr. Med. Chem. 13* (2006) 697–710.
13. D.B. Rusch, A.L. Halpern, G. Sutton, K.B. Heidelberg, S. Williamson, S. Yooseph, D. Wu, J.A. Eisen, J.M. Hoffman, K. Remington, *et al.*, The Sorcerer II Global Ocean Sampling expedition: Northwest Atlantic through eastern tropical Pacific, *PLoS Biol. 5* (2007) e77.
14. Bioserv server, Faculty of Food Technology and Biotechnology, University of Zagreb, Zagreb, Croatia (*http://bioserv.pbf.hr/*).
15. S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, *J. Mol. Biol. 215* (1990) 403–410.
16. M.Z. Ansari, G. Yadav, R.S. Gokhale, D. Mohanty, NRPS--PKS: A knowledge-based resource for analysis of NRPS/PKS megasynthases, *Nucleic Acids Res. 32* (2004) W405–W413.
17. H. Tae, E.B. Kong, K. Park, ASMPKS: An analysis system for modular polyketide synthases, *BMC Bioinformatics, 8* (2007) 327.
18. J. Cullum, A. Starcevic, J. Zucko, M. Elbekali, N. Skunca, D. Kovacek, J. Diminic, V. Zeljeznak, D. Pavlinusic, J. Simunkovic, P.F. Long, D. Hranueli, *In silico* generation of novel polyketides for the use in agro-industry, *Proceedings of the 2008 Joint Central European Congress: The 4th Central European Congress on Food and the 6th Croatian Congress of Food Technologists, Biotechnologists and Nutritionists, Vol. 2*, D. Ćurić (Ed.), Zagreb, Croatia (2008) pp. 287–294.
19. A. Starcevic, J. Zucko, J. Simunkovic, P.F. Long, J. Cullum, D. Hranueli, *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* (2008) 6882–6892.
20. A.L. Delcher, K.A. Bratke, E.C. Powers, S.L. Salzberg, Identifying bacterial genes and endosymbiont DNA with Glimmer, *Bioinformatics, 23* (2007) 673–679.
21. J. Besemer, M. Borodovsky, GeneMark: Web software for gene finding in prokaryotes, eukaryotes and viruses, *Nucleic Acids Res. 33* (2005) W451–454.
22. P. Rice, I. Longden, A. Bleasby, EMBOSS: The European Molecular Biology Open Software Suite, *Trends Genet. 16* (2000) 276–277.
23. S.R. Eddy, Profile hidden Markov models, *Bioinformatics, 14* (1998) 755–763.
24. A. Bateman, E. Birney, L. Cerruti, R. Durbin, L. Etwiller, S.R. Eddy, S. Griffiths-Jones, K.L. Howe, M. Marshall, E.L. Sonnhammer, The Pfam protein families database, *Nucleic Acids Res. 30* (2002) 276–280.
25. TMSS web portal, Thiotemplate Modular Systems Studies (*http://bioserv.pbf.hr/cms/*).
26. J. Rumbaugh, I. Jacobson, G. Booch: *The Unified Modeling Language Reference Manual*, Addison-Wesley, Boston, MA, USA (2005).
27. BioSQL schema, Open Bioinformatics Foundation, USA (*http://obda.open-bio.org/*).
28. Perl programming language, Perl Foundation, Gran Ledge, MI, USA (*http:// www.perl.org/*).
29. The Source for Java Developers, Oracle and Sun, Redwood Shores, CA, USA (*http://java.sun.com/*).
30. The JavaScript Source, Jupitermedia Corp., Darien, CT, USA (*http:// javascript.internet.com/*).

31. D. Weininger, SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules, *J. Chem. Inf. Comput. Sci. 28* (1988) 31–36.

32. Jmol: An open-source Java viewer for chemical structures in 3D (*http://www.jmol.org/*).

33. F. Csizmadia, JChem: Java applets and modules supporting chemical database handling from web browsers, *J. Chem. Inf. Comput. Sci. 40* (2000) 323–324.

34. WWW READSEQ Sequence Conversion, National Institutes of Health, Bethesda, MA, USA (*http://www-bimas.cit.nih.gov/molbio/readseq/*).

35. Google Web Toolkit. Chapter 10 Database Editor on Google App Engine Java, USA (*http://www.gwtapps.com/*).

36. M. Oliynyk, M. Samborskyy, J.B. Lester, T. Mironenko, N. Scott, S. Dickens, S.F. Haydock, P.F. Leadlay, Complete genome sequence of the erythromycin-producing bacterium *Saccharopolyspora erythraea* NRRL23338, *Nat. Biotechnol. 25* (2007) 447–453.

37. The genome of *Streptomyces scabies* strain 87.22, Wellcome Trust Sanger Institute, Hinxton, UK (*http://www.sanger.ac.uk/Projects/S_scabies/*).

38. J. Zucko, N. Skunca, T. Curk, B. Zupan, P.F. Long, J. Cullum, R. Kessin, D. Hranueli, Polyketide synthase genes and the natural products potential of *Dictyostelium discoideum*, *Bioinformatics, 23* (2007) 2543–2549.

39. A. Starcevic, S. Akthar, W.C. Dunlap, J.M. Shick, D. Hranueli, J. Cullum, P.F. Long, Enzymes of the shikimic acid pathway encoded in the genome of a basal metazoan, *Nematostella vectensis*, have microbial origins, *Proc. Natl. Acad. Sci. USA, 105* (2008) 2533–2537.

40. R.J, Zauhar, G. Moyna, L. Tian, Z. Li, W.J. Welsh, Shape signatures: A new approach to computer-aided ligand- and receptor-based drug design, *J. Med. Chem. 46* (2003) 5674–5690.

41. P. Shen, H.V. Huang, Homologous recombination in *Escherichia coli*: Dependence on substrate length and homology, *Genetics, 112* (1986) 441–457.

42. B. Springer, P. Sander, L. Sedlacek, W.-D. Hardt, V. Mizrahi, P. Schär, E.C. Böttger, Lack of mismatch correction facilitates genome evolution in mycobacteria, *Mol. Microbiol. 53* (2004) 1601–1609.

43. C. Rayssiguier, D.S. Thaler, M. Radman, The barrier to recombination between *Escherichia coli* and *Salmonella typhimurium* is disrupted in mismatch-repair mutants, *Nature, 342* (1989) 396–401.

44. D. Denapaite, A. Paravić Radičević, B. Čajavec, I.S. Hunter, D. Hranueli, J. Cullum, Persistence of the chromosome end regions at low copy number in mutant strains of *Streptomyces rimosus* and *S. lividans*, *Food Technol. Biotechnol. 43* (2005) 9–17.

45. H. Petković, J. Cullum, D. Hranueli, I.S. Hunter, N. Perić-Concha, J. Pigac, A. Thamchaipenet, D. Vujaklija, P.F. Long, Genetics of *Streptomyces rimosus*, the oxytetracycline producer, *Microbiol. Mol. Biol. Rev. 70* (2006) 704–728.

46. J.C. Alonso, G. Lüder, T.A. Trautner, Intramolecular homologous recombination in *Bacillus subtilis* 168, *Mol. Gen. Genet. 236* (1992) 60–64.

47. M.S. Dillingham, M. Spies, S.C. Kowalczykowski, RecBCD enzyme is a bipolar DNA helicase, *Nature, 423* (2003) 893–897.

48. G. Yadav, R.S. Gokhale, D. Mohanty, SEARCHPKS: A program for detection and analysis of polyketide synthase domains, *Nucleic Acids Res. 31* (2003) 3654–3658.

49. E. Zazopoulos, K. Huang, A. Staffa, W. Liu, B.O. Bachmann, K. Nonaka, J. Ahlert, J.S. Thorson, B. Shen, C.M. Farnet, A genomics-guided approach for discovering and expressing cryptic metabolic pathways, *Nat. Biotechnol. 21* (2003) 187–190.

50. S.B. Zotchev, A.V. Stepanchikova, A.P. Sergeyko, B.N. Sobolev, D.A. Filimonov, V.V. Poroikov, Rational design of macrolides by virtual screening of combinatorial libraries generated through *in silico* manipulation of polyketide synthases, *J. Med. Chem. 49* (2006) 2077–2087.

51. C. Rausch, T. Weber, O. Kohlbacher, W. Wohlleben, D.H. Huson, Specificity prediction of adenylation domains in nonribosomal peptide synthetases (NRPS) using transductive support vector machines (TSVMs), *Nucleic Acids Res. 33* (2005) 5799–5808.

52. H. Jenke-Kodama, E. Dittmann, Evolution of metabolic diversity: Insights from microbial polyketide synthases, *Phytochemistry, 70* (2009) 1858–1866.

53. T. Weber, C. Rausch, P. Lopez, I. Hoof, V. Gaykova, D.H. Huson, W. Wohlleben, CLUSEAN: A computer-based framework for the automated analysis of bacterial secondary metabolite biosynthetic gene clusters, *J. Bacteriol. 140* (2009) 13–17.

54. M.H.T. Li, P.M.U. Ung, J. Zajkowski, S. Garneau-Tsodikova, D.H. Sherman, Automated genome mining for natural products, *BMC Bioinformatics*, *10* (2009) 185.