In silico analysis of polyketide synthases

Dem Fachbereich Biologie der Technischen Universität Kaiserslautern
zur Erlangung des akademischen
Grades "Doktor der Naturwissenschaften" eingereichte

DISSERTATION

Vorlegt von Jurica Žučko

Vorsitzender: Prof. Dr. Matthias Hahn
Erster Berichterstatter: Prof. Dr. John Cullum
Zweiter Berichterstatter: Prof. Dr. Regine Hakenbeck

Lehrbereich Genetik der Technischen Universität Kaiserslautern, Juli 2010

D386

But apart from the sanitation, the medicine,
education, wine, public order, irrigation, roads,
the fresh-water system, and public health, what
have the Romans ever done for us?
Monty Python's Life of Brian

PUBLISHED SCIENTIFIC PAPERS

The thesis is a cumulative thesis based on the following four published papers. The papers bound in the thesis are identical to the published versions.

Pavle Goldstein, **Jurica Zucko**, Dušica Vujaklija, Anita Krisko, Daslav Hranueli, Paul F. Long, Catherine Etchebest, Bojan Basrak and John Cullum. **Clustering of protein domains for functional and evolutionary studies.** *BMC Bioinformatics*, ***10***, 335, 2009.

Antonio Starcevic, A., **Jurica Zucko**, Jurica Simunkovic, Paul F. Long, John Cullum and Daslav 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***, 6882-6892, 2008.

Gaetano Castaldo*, **Jurica Zucko***, Sibylle Heidelberger, Dušica Vujaklija, Daslav Hranueli, John Cullum, Pakorn Wattana-Amorn, Matthew P. Crump, John Crosby and Paul F. Long. **Proposed arrangement of proteins forming a bacterial type II polyketide synthase.** *Chem. Biol.*, ***15***, 1156-1165, 2008.

Jurica Zucko, Nives Skunca, Tomaz Curk, Blaz Zupan, Paul F. Long, John Cullum, Richard Kessin and Daslav Hranueli. **Polyketide synthase genes and the natural products potential of *Dictyostelium discoideum*.** *Bioinformatics*, ***23***, 2543–2549, 2007.

* These authors contributed equally to this work

Table of Contents

Introduction

1. Introduction

Since the discovery of penicillin natural products have been extensively used in human medicine and have helped to improve the quality of life. They have contributed to the greatly extended life span in the last century. Even today, natural products remain the most important "weapons" against diseases with more than 60 % of approved pharmaceutical products being either natural products or their derivatives. (Demain, 2009). The interest of the pharmaceutical industry for natural products is not fading as natural products and their derivatives make up more than 50 % of newly introduced drugs into the market in the last 25 years (Newman and Cragg, 2007). Through the time various sources of natural products have been used. In the past plants have been the major sources of them. In the middle of the last century microorganisms became the major source of active compounds from nature. Especially rich in secondary metabolites are bacteria from the soil-dwelling *Streptomyces* genus with various species being widely used as producers for pharmaceutical industry (Bentley *et al*., 2002; Ikeda *et al*., 2003). In the last decades other niches have been explored and the one showing most promise in richness and diversity of life, as well as possible natural products, are oceans (Newman and Cragg, 2004; Dunlap, *et al*., 2006) - as it was shown in recent metagenomic study of Atlantic and Pacific oceans (Rusch *et al*., 2007). Usually natural products are secondary metabolites which help their producers either to "cope better" with the environment or to gain certain advantages over other organisms in the habitat. Large numbers of these compounds show antimicrobial properties and have been extensively employed in medicine. Today natural products and their derivatives make up almost 80 % of antibacterial drugs used (e.g. erythromycin, tetracyclins, rifamycins). Emerging fields of usage are as anticancer drugs (e.g. epothilones, doxorubicin, daunorubicin), where natural products are the basis for 74 % of all new chemical entities. They are also helping in diseases specific for modern society, like high blood pressure (e.g. captopril-ACE inhibitor) and high cholesterol levels (e.g. lovastatin), in which natural products constitute more than half of the compounds used (Demain, 2009).

With the growing need for new biologically active compounds to help humans cope with the ailments typical of modern society, the growing number of resistant pathogens and the advances in technology which allows us to obtain genetic makeup of a large number of

microorganisms, a lot of approaches and methods have been tried to obtain novel biologically active compounds. Some methods focus on more efficient ways of screening natural habitats for active compounds (Shen, *et al*., 2003; Foerstner and Bork 2007; Shu 1998; Singh and Pelaez, 2008; Van Lanen and Shen, 2006) while others try to modify or rearrange already know "factories" of natural products (Hranueli *et al*., 2005; Menzella *et al*., 2005; Cane *et al*., 1998; Sherman, 2005; Bachmann, 2005). In this thesis an attempt has been made to improve recognition and specificity of an enzyme family that synthesizes one chemical class of the natural products – the polyketides.

1.1 Polyketide synthases

Polyketides are a diverse class of chemical compounds synthesized by the secondary metabolism of bacteria, fungi, plants and animals. They have a wide range of pharmaceutical properties, including antibacterial (e.g. erythromycin, rifamycin B), anticancer (e.g. epothilone), anticholesterol (e.g. lovastatin) and immunosuppressant (e.g. rapamycin) (Staunton and Weissman, 2001). Polyketides are synthesized by a family of enzymes called polyketide synthases, (PKS) [similar to fatty acid synthases (FAS)], which differ in organisation and structure of the enzyme complex but all have a common pattern of biosynthesis. No matter the type of the enzyme used, they are all assembled by successive rounds of decarboxylative Claisen condensations between a thioesterified acyl extender unit and a growing acyl thioester chain. As building blocks residues of acetate (malonyl-CoA) and propionate (methylmalonyl-CoA) are mostly used, but more complex units can also be used (such as benzoyl-CoA, 3,4-dihydroxycyclo-hexanecarbonyl-CoA (3,4-DHCHC-CoA), 3-Methylbutyryl-CoA (Chan *et al*., 2009)

Every unit contributes with two carbon atoms to the assembly of the linear chain (backbone) and the β-carbon always carries a keto group which can be reduced to hydroxyl or fully removed by dehydration and enoyl reduction during the biosynthesis. This alternating occurrence of keto groups is responsible for the name "polyketide" for this group of compounds. The other (α) carbon atom incorporated in the backbone chain by each building unit can carry different substituents depending on the building block used (H for acetate, CH_3 for propionate, CH_3CH_2 for butyrate etc.). Although having a common mechanism of

biosynthesis polyketides are a very diverse class of compounds. Mechanisms that contribute to great diversity of polyketide products are incorporation of various substrates, different reduction level of keto groups on the β-carbon atom, possibility of various chiral configurations of the branching groups and the total length of the chain synthesized (Hopwood and Sherman 1990).

1.1.1 Type I polyketide synthases

As mentioned before, PKSs are classified in three groups (Fig. 1) based on the organisation of the enzyme complex. The best understood are type I PKSs, which are further divided in two subgroups - iterative and modular. Organisationally they resemble animal FAS - with same type and order of catalytic domains. The difference is in the presence/activity of reduction domains, which are not obligatory in PKSs, and in the specificity of acyltransferase domains (AT), which in PKSs can select several substrates while in fatty acid biosynthesis only malonyl-CoA is used. The main characteristic of type I PKSs is the presence of several functional domains on a single polypeptide, which is organised in the higher organisational structure called module. A module carries out one elongation cycle of the polyketide product.

Iterative type I PKS consist of only one set of domains (one module) which are used iteratively until the desired chain length is reached. They are, therefore pre-programmed. A minimal set of domains needed for the biosynthesis consists of ketosynthase (KS), acyltransferase (AT) and acyl carrier protein (ACP) domain. Optional domains that can be part of the module are ketoreductase (KR), dehydrogenase (DH) and enoyl reductase (ER) and are involved in processing of the beta hydroxyl group of incorporated building unit. The mechanism of biosynthesis of these compounds follows the general mechanism of polyketide biosynthesis but the regulatory mechanisms are still not fully understood as the same enzyme complex is capable of using different substrates and carrying out different levels of reduction processing at different iteration cycles (Staunton and Weissman 2001).

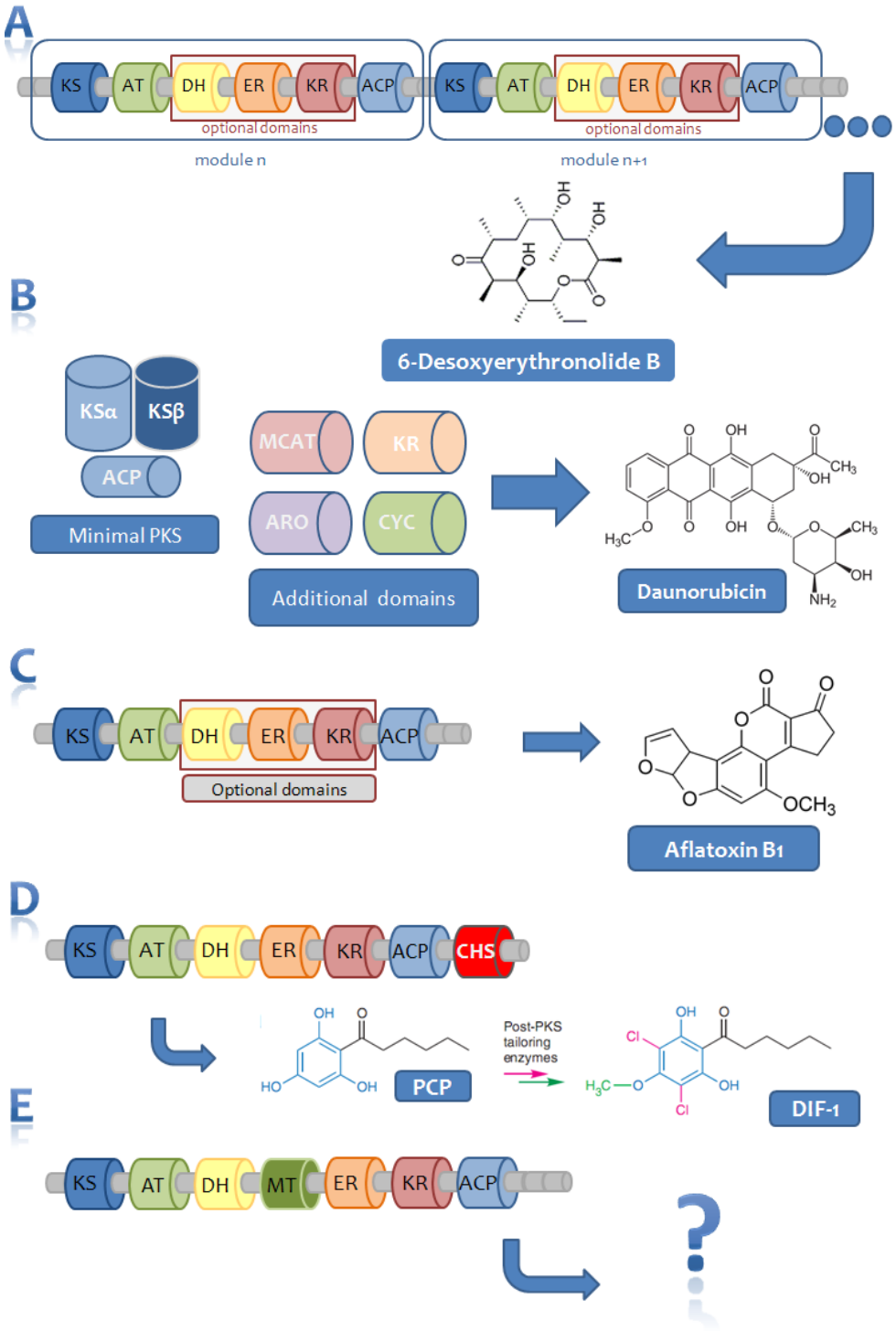

Fig. 1 Examples of typical organisations of several groups of PKSs and their respective products. **A** – schematic representation of modular type I PKS and its model compound - erythromycin. The 6-desoxyerythronolide B (6DEB) – a precursor of the erythromycin is shown. **B** – Type II PKS. The PKS involved in the biosynthesis of daunorubicin/doxorubicin and one product are shown. **C** - the schematic representation of the iterative type I PKS typical for fungi and bacteria with the typical example of the synthesized product – aflatoxin. **D** – the schematic representation of the PKSI-PKSIII hybrid from *Dictyostelium discoideum* with the characterised product. **E** – the schematic representation of the typical *D. discoideum* PKS gene.

Studies of the typical example for this type of enzyme, lovastatin PKS, suggested that other enzymes involved in the biosynthetic pathway may play a role in modulating the overall activity of the synthase, thus granting it the ability to discriminate between polyketide intermediates at different stages of assembly (Kennedy *et al*., 1999).

Much better understood are the modular type I PKSs, and their model system – erythromycin PKS. Modular type I PKS consists of several sets of domains organised in modules on a single polypeptide. Each module is responsible for catalysis of one complete cycle of polyketide elongation and associated functional group modifications. Two types of modules can be distinguished – extender modules and starter or loading modules. Each extender module must minimally contain ketosynthase (KS), acyltransferase (AT) and acyl carrier protein (ACP) domain. Optional are domains involved in processing the beta hydroxyl group of incorporated substrate: ketoreductase (KR), dehydrogenase (DH) and enoyl reductase (ER). The starter module does not have the minimal set of domains and varies greatly between different PKSs. The process of polyketide biosynthesis occurs in four phases: priming of the apo-synthase, chain initiation, chain elongation, and termination. In this first step, *priming of the apo-synthases,* conversion of inactive apo-form of acyl carrier protein (ACP) to the active, phosphopantetheinyl containing, holo-ACP form is taking place. After the conversion process each ACP carries a 20 Å–long phosphopantetheinyl prosthetic group. The phosphopantetheinyl prosthetic group ends in a cysteamine thiol group that becomes the site of covalent attachment for acyl groups and serves as a flexible tether for both the monomer units and growing polyketide chain. The ACP is, after the conversion to its active holo form, able to support initiation, elongation and termination of its characteristic polyketide product (Cane *et al*., 1998).

Type I modular PKSs are considered to be "genetically pre-programmed assembly lines" – as it is possible in principle to predict the chain length, building units used and level of reduction of each building unit of the polyketide product from the DNA sequence of respective polyketide synthase.Recently PKSs capable of skipping or re-using a module have been discovered. Although the mechanisms of these actions are still unclear, it shows that synthases can somehow bypass their "genetic programming", thus, giving even more potential to increase diversity of synthesized products (Moss *et al*., 2004; Wenzel and

Müller, 2005). In the next step of biosynthesis, the initiation of the polyketide chain, an activated monomer (acyl-CoA thioester) is loaded onto the first holo-ACP. The resultant acylthio-enzyme intermediates then act as a donor for the first elongation step. Modules involved in the initiation step of polyketide biosynthesis are one of the major sources for polyketide diversity as most of the unusual building units are incorporated by loading modules, e.g. isobutyryl-CoA (avermectin), cyclohexenoyl-CoA (FK506), and 3-amino-5-hydroxybenzoyl-CoA (rifamycin). After the initiation of the biosynthesis, there follow a series of elongation cycles (Cane *et al*., 1998).

For carrying out the elongation step a minimal module (KS, AT and ACP) and an upstream ACP domain supplying the donor chain are required. The mechanism of each elongation cycle is as follows: the upstream acyl group (growing polyketide chain) is transferred from the ACP onto an active site cystein of the KS domain in the downstream (next) module. The KS then carries out the decarboxylation of the acyl-S-ACP [acyl = (methyl)malonyl] generating a carbanion which elongates and translocates polyketide chain from the KS domain to the ACP domain in the same module. The newly generated polyketide chain, which has been extended by two carbon atoms in the backbone, then serves as a donor for the next module. If additional (reduction) domains are present in the module modifications related to the β-carbon atom can occur. The presence of all reduction domains causes a full reduction of β-ketoacyl intermediate to the saturated methylene. Inactivation or absence of one or more of the reduction domains (KR, DH and ER) causes no or partial reduction of the product (β-keto, β-hydroxy, or α,β-unsaturated) (Keatinge and Walsh, 1999). When the product reaches the final module of the synthase polyketide chain, termination with release of the full-length polyketide chain occurs. The chain can be released from the synthase directly by hydrolysis, giving rise to the free acid, or by intramolecular capture by a hydroxyl of the acyl chain itself, giving rise to a lactone product. The other possibility for chain termination and release, not present in all PKSs, is the use of a terminal thioesterase domain (TE). The thioesterase has an active serine site to which the acyl chain can be transferred from the last ACP domain. The acyl-TE intermediate polyketide chain is again cleaved either by hydrolysis or by cyclization (Cane *et al*., 1998). When the polyketide chain is fully synthesized and released from the PKS it usually undergoes further enzymatic tailoring by auxiliary enzymes. Most of the enzymes involved in post-PKS modifications are specific to

the PKS pathway itself and are usually found as part of the PKS cluster. Common activities used in post-PKS modification are hydroxylases, glycosyl transferases and methyl transferases, which for example in erythromycin PKS transform biologically inactive 6-DEB, released from the synthase, to erythromycin A, a widely used antibiotic (Staunton and Weissman, 2001).

1.1.2 Type II polyketide synthases

Unlike type I, which form large multienzyme complexes, type II PKSs are composed of several individual, monofunctional or bifunctional enzymes that form a dissociable complex. Again, several proteins form a so called "minimal PKS" which consists of two ketosynthase units - called KS_α and KS_β and an acyl carrier protein (ACP). Additional subunits can include ketoreductases, cyclases, aromatases, oxygenases, as well as gylcosyl and methyl transferases. The minimal PKS catalyses the iterative decarboxylative condensation of malonyl-CoA extender units with an acyl starter unit (Hertweck *et al*., 2007). Both KS subunits have a high sequence identity and, as was found from recent structural studies of the PKS type II complex, have evolved highly complementary contacts (Keatinge-Clay *et al.,* 2004). The crucial difference between the subunits is the lack of the active site cystein in the KS_β subunit. Therefore, based on analogy with other types of PKSs, the KS_α subunit obviously catalyzes the Claisen-type condensation between the nascent polyketide chain and malonyl units. The function of the KS_β subunit is still rather unclear. Its main role is determining the length of the synthesized polyketide chain and because of this property it is sometimes also called the chain length factor (CLF). Recently the idea prevailed that the chain length is determined by "measuring" its length. It is "measured" in a protein cavity located at the interface of the heterodimer (KS_α and KS_β subunit) and several residues in KS_β, named gatekeepers, involved in determining the length of the cavity (channel) were identified (Burson and Khosla, 2000; Tang *et al*., 2003; Keatinge-Clay *et al*., 2004). Although KS_β is the primary determinant of polyketide chain length, it is definitely not the only one. It was shown that cyclases can influence the length of the chain (Shen *et al*., 1999) and is currently speculated that chain length is, in some limited degree, determined by entire PKS complex. It has been proposed that the KS_β subunit also has catalytic activity. It was found that it is involved in loading of malonyl-CoA and generating acetyl by decarboxylation of malonyl-ACP

in the actinorhodin and tetracenomycin PKSs (Bisang *et al*., 1999). The other candidate for supplying malonyl building units is a malonyl-CoA:ACP transferase (MCAT). However, as MCAT is missing in the majority of PKS type II clusters and was experimentally shown not to be required for *in vitro* polyketide synthesis (Matharu *et al*., 1998) it was not included in the type II minimal PKS. Adding more confusion to the debate about which protein supplies the malonyl building units was the discovery of Simpson and collaborators (Arthur *et al*., 2006) who found that ACP is capable of self-malonylation *in vitro*. Unlike type I PKSs, all known type II PKSs use only malonyl-CoA as an extender unit but which protein is responsible for supplying it is still remains unclear. After the polyketide product is fully synthesized it can undergo ketoreduction, if a KR domain is present in the synthase, followed by cyclization and aromatisation of the polyketide product.

The type II PKS producing the anticancer drug daunorubicin and its derivative doxorubicin, from *S. peucetius* (Grimm *et al*., 1995), was used as a model for the investigation of protein interactions within the complex and for the overall quaternary structure of the complex. The information about the overall three dimensional (3D) organisation of the type II PKS complexes is very limited, as the only crystal structure involving more than one discrete protein is that of KS/CLF heterodimer involved in the biosynthesis of actinorhodin in *S. coelicolor* (Keatinge-Clay *et al*., 2004). Much more structural information is available for the individual proteins of the polyketide type II synthases. The protein with the most 3D structures available is the ketosynthase (KS_α) domain, with more than 25 structures having 40 % sequence identity to the ketosynthase of the actinorhodin PKS (data accessed from PDB database on March 3, 2010 http://www.pdb.org/pdb). Its homologue, the chain length factor (CLF) has a similar number of homologue structures in the Protein Data bank (PDB) but with lower sequence identity (> 30 %). For the last part of the minimal synthase, the ACP domain, several structures from similar systems are available (Crump *et al*., 1997; Li *et al*., 2003). Structural data from related organisms was also available for most of the proteins which are not part of the minimal synthase complex, namely KR (Hadfield *et al*., 2004) and MCAT (Keatinge-Clay *et al*., 2003) from the *S. coelicolor* actinorhodin PKS, cyclase from *S. nogalater* (Sultana *et al*., 2004) and *S. glaucescens* (Thomspon *et al*., 2004) and aromatase/cyclase from *S. glaucescens* (Ames *et al*., 2008).

1.1.3 Type III polyketide synthases

The type III PKSs are often called chalcone/stilbene synthase (CHS/STS). They were first identified in plants, as these enzymes are catalyze the first steps of the flavonoid biosynthesis pathway. Later they were also found in bacterial genomes but were regarded as more typical of plant secondary metabolism. Unlike other two types of PKSs, the type III PKSs are relative small in size (350-390 amino acids). They also show a significant difference to other types of PKSs, both in organization and functioning of the enzyme. The type III PKSs are quite a diverged protein family, most of them sharing about 25 % amino acid sequence identity with CHS and with each other. Within the type III PKS protein family of enzymes there are differences in the preference for starter molecule, the number of acetyl additions (iterations) they catalyze, the mechanism of chain termination and the pattern of intermolecular cyclization. The best studied member of the family is chalcone synthase (CHS) which functions in a homodimer form. It contains an acyltransferase activity which loads the p-coumaroyl-CoA starter unit onto a catalytic cysteine, a decarboxylase activity that activates malonyl-CoA, an iterative condensing activity that couples the resulting acetyl anion to the growing ketide chain, a cyclase activity that forms the cyclised polyketide precursor of chalcone *via* an intermolecular Claissen condensation of the linear tetraketide intermediate, and an aromatase-like activity at the same catalytic site. Using the CHS crystal structure core chemical machinery of the type III PKS, the active site was identified as a catalytic trio (Cys, His, Asn) positioned at the top of the active site cavity (Austin and Noel, 2003).

The *Dictyostelium* genome revealed two type III PKS which are unique in enzyme organisation as both occur as C-terminal fusions to multidomain polypeptides homologous to type I PKS/FAS. It is believed that the acylthioester product of the type I PKS part of the enzyme is directly transferred from the pantetheine arm of the ACP domain to the catalytic cysteine of the neighboring type III PKS domain and serves as a starter unit. The crystal structure of the type III PKS domain shows it has a homodimeric structure and the active site cavity contains the same catalytic triad previously observed in all type III PKSs with known protein structure (Austin *et al*., 2006).

1.1.4 Specificity of type I PKS domains

Polyketide compounds represent one of the biggest reservoir of active compounds used in medicine (Demain, 2009). Due to the chronic lack of new drugs, especially antibiotics, pharmaceutical companies have started to develop methods of generating novel polyketide compounds from known clusters. This is being done in several ways:

- by changing the number of modules (genes) in PKSs which causes changes in polyketide chain length,
- by changing/substituting the acyl transferase (AT) domain which selects the building unit incorporated into the polyketide chain,
- by changing (e.g. deleting) the reduction domains from the module thus altering the level of reduction, and finally
- by changing the stereochemistry at centres which carry alkyl and hydroxyl groups by altering the domains determining it.

The first attempts to generate novel polyketide compounds started in the middle of the 90's by deleting end modules of the erythromycin synthase and moving the final thioesterase (TE) domain to upstream modules (Cortes *et al*., 1995). Soon after followed experiments in which one or more modules at the beginning of the synthase were deleted and the PKS supplied with various intermediary products (Jacobsen *et al*., 1997). In both cases the newly created synthases synthesised the expected/predicted products, although at much lower yields than the "original" (unaltered) synthase but the experiments showed a certain amount of robustness and tolerance "incorporated" in the synthase itself which increases the opportunity of creating novel hybrid complexes. After changes affecting the entire module the more subtle alterations were carried out on the level of the modules themselves by replacing entire domains with domains from other modules/PKSs (Oliynyk *et al*., 1996; Bedford *et al*., 1996). Swapping of AT domains as well as KR domains showed that most of the domains are quite robust in accepting intermediary products and carrying out the expected enzymatic reactions. It was also shown that there is a certain degree of dependency/correlation both between domains in a specific module as well as between modules themselves for synthases to function at an optimal level. These examples gave a new boost to the polyketide genetic/enzyme engineering field at the end of the 90's and

studies started researching on even smaller subunits of the PKS – a single domain, its specificity and factors determining it. The first research dealing with the question of substrate-determining residues in AT domains was done by Peter Leadlay's group (Haydock *et al*., 1995). Based on alignment of AT domains specific for malonyl-CoA and methylmalonyl-CoA, they identified variable regions within the AT domain which allowed them to unambiguously assign newly-sequenced domains to a specific subgroup. The group of Chaitan Khosla used experimental methods to identify a short variable C-terminal segment of the AT domain as the principal determinant of substrate specificity (Lau *et al*., 1999). The KOSAN Biosciences*,* Inc*.* used cassette replacement of AT domains in DEBS with heterologous AT domains with different substrate specificities to generate a library of polyketide compounds. As several replacements of the AT domain in module 4 of DEBS had failed, site-directed mutagenesis of specific residues believed to be involved in determining substrate specificity was successfully carried out. This was the first example in which substrate specificity of an extender PKS module has been altered using site-specific mutagenesis (Reeves *et al*., 2001).

Two later studies used only computational tools to detect PKS domains present in protein sequences and then determine the specificity for each AT domain. Yadav and collaborators (Yadav *et al*., 2003) identified domains of modular PKSs based on their sequence similarity. As a method of determining similarity BLAST (Altschul *et al*., 1990) was used and domains from erythromycin synthase module 4 were used as queries for each domain type. The method gave in general good recognition of PKS domains (90-100 % accuracy) with most problems caused by the reduction domains, in particular the dehydratase (DH) domains, as well as the acyl carrier protein (ACP) domains. For prediction of the substrate specificity of AT domains, two fingerprints of the active site residues, which are believed to be involved in determining substrate specificity, were created. The fingerprint can identify AT domains specific for the two most common substrates, malonyl-CoA and methylmalonyl-CoA, with high probability. These fingerprints are an extension of previously identified specificity determining residues (Haydock *et al*., 1995; Lau *et al*., 1999) and were defined from a much larger sample, giving them much greater robustness and precision. The fingerprints were extracted from the alignment of PKS AT domains with the AT domain of *Escherichia coli* FAS in which residues involved in determining specificity were identified from 3D structure of the

respective *E. coli* AT domain (Serre *et al*., 1995). In the other paper dealing with computational analysis of polyketide sythases Minowa and collaborators (Minowa *et al*., 2007) used similar methods. They identified PKS domains using homology search with Hidden Markov Model (HMM) profiles (Eddy, 1998) which is a more sensitive method than BLAST used by Yadav and co-workers (Yadav *et al*., 2003). From the multiple alignment of AT domains substrate subgroups were defined based on literature data. Conserved sites within the specific subgroup were extracted and from extracted residues HMM profiles were built using HMMER (Eddy, 1998). These profiles, containing only residues conserved within specific subgroup of AT domain, were then used to search all domains whose substrate specificity was unknown. In total 13 HMM profiles able to distinguish substrate specificity of AT domain were created and had 95 % correct predictions on a test sample.

Other types of domains which contribute to the variability of synthesized polyketide are reductive domains, of which the KR domain, which reduces the keto group and then determines the stereochemical configuration of resulting hydroxyl group, is the best characterized one. Reid and collaborators (Reid *et al*., 2003) experimentally determined residues of the catalytic triad involved in reduction of the keto group and correlated some residues with the stereochemical outcome of the keto reduction. Later on Caffrey (Caffrey, 2003) identified several conserved residues indicative for determining stereochemical outcome of the keto reduction based on the multiple alignment of both types of KR domains (-OH R and S configuration). Based on molecular modelling it was suggested that the KR domain also plays a role in determining the stereochemistry of methyl group (Starcevic *et al*., 2007) if propionate is incorporated into the polyketide chain. After the crystal structure of erythromycin ketoreductase (KR) domain was solved (Keatinge-Clay and Stroud, 2006) residues determining both stereochemistry of hydroxyl group as well as the stereochemistry of an α-substituent were identified. Based on those residues six motifs (fingerprints) can be created covering all possible outcomes of ketoreductase and epimerase activity. The dehydratase (DH) domain, which catalyzes dehydration and thus creates double bond between C_β and C_α atoms, has after publication of its crystal structure again became a subject of interest. Previous studies on the prediction of activity of DH domains based on conserved motifs from a limited number of clusters were not applicable to all DH domains and often failed due to the high variability of DH domains (Tang *et al*., 1998). Based on

multiple alignment of DH domains from several clusters and the structure of erythromycin module 4 DH domain, Keatinge-Clay (Keatinge-Clay, 2008) identified several conserved motifs. Two of them are associated with catalytic residues and can be used to determine whether a DH domain is active. One is associated with the putative epimerisation function of the DH domain and one is thought to be involved in docking of the ACP domain. After dehydration the resulting polyketide can have either *cis* or *trans* conformation which was believed to be determined by DH domain. Analysis of both "*cis*" and "*trans*" DH domains failed to identify residues which might contribute to the resulting conformation. The current hypothesis is that the primary determinant whether a *cis* or *trans* double bond will be formed is the cooperating KR domain, i.e. the stereochemistry of the β-hydroxyl group (Keatinge-Clay, 2008). Although the structure of DH domain has been solved its exact functions, specificities and interactions are still surrounded by a shroud of mystery and need further investigations.

The enoyl reductase (ER) domain is the last of the reduction domains regarding the timing of its action in the process of polyketide biosynthesis. It catalyzes reduction of the double bond (enoyl group) created by DH domain to an alkyl group. The mechanism of PKS ER reaction was deduced from FAS studies and involves 1,4-nucleophilic addition of the hydride ion from the coenzyme NADPH to the unsaturated thioester intermediate, followed by stereospecific protonation at the α-carbon. If propionate, or some other building unit with an alkyl group on the α-carbon is used as the extender unit, reduction carried out by ER domain determines the configuration of the Cα methyl (alkyl) group. Recently a tyrosine residue in the ER active site has been correlated to the chirality of the methyl branch that is introduced (Kwan *et al*., 2008). The occurrence of the tyrosine residue implies the S configuration of the methyl group, its absence implies R configuration. Mutagenesis of this residue caused a switch from S to R configuration of the methyl group, but not in all cases which suggests some additional residues might be involved in determining the chirality of methyl group. Regarding the activity of the ER domain, up till now the motif used as a NADPH binding site has been identified and its mutagenesis caused ER to become inactive (Witkowski *et al*., 2004). However, the residues forming the active site have not yet been identified until now neither in PKSs nor in the animal FASs (Smith and Tsai, 2007). Experiments exploring substrate specificity of ER domains showed they exhibit relatively relaxed substrate specificity and that

ER domains can usually be freely swapped between various polyketide synthases (Khosla *et al*., 1999). Future experiments should show which approach is more profitable - domain swapping or site directed mutagenesis to obtain novel polyketides by engineering of the ER domain.

1.1.5 Computational analysis of modular biosynthetic clusters

Several platforms for the detection and analysis of modular biosynthetic cluster already exist. The *de facto* standard for the analysis of modular polyketide synthases is the SEARCHPKS program (Yadav *et al*., 2003). It does not have gene finding tools implemented in the program so it requires protein sequence as an input. Protein sequence is then searched using BLAST (Altschul *et al*., 1997) with PKS domains from erythromycin synthase used as queries. The only domain undergoing further analysis is the aclytransferase (AT) domain. Substrate specificity of AT domain is determined by extracting 13 specificity determining residues (Yadav *et al*., 2003) from its alignment with the crystal structure of AT domain from *Escherichia coli* FAS (Serre *et al*., 1995). Extracted motif is then compared with motifs from AT domains with known substrate specificity. If there is an identical match query the AT domain is assigned the same specificity as that of the matched AT domain. If an identical match cannot be found the program shows to the user extracted motif and motifs specific for known substrate. The program has motifs specific only for malonate and methylmalonate substrates. When all analyses are finished the program shows graphical representation of the entire cluster with domains, modules and linkers easily distinguishable and for each module it shows chemical structure incorporated by the module into polyketide. The chemical structure shows the reduction level for each block based on presence of reduction domains in the module but does not show which substrate is incorporated by the module.

The other popular program used for analysis of modular polyketide synthases is the MAPSI system (Tae *et al*., 2009), the successor of ASMPKS system (Tae *et al*., 2007). It searches microbial genomes for modular PKS clusters and detects them using a similarity search implemented with BLAST (Altschul *et al*., 1997) against an integrated database of annotated polyketides. If MAPSI is unable to detect homologous PKS clusters in the genome sequence it

searches the remaining proteins with HMM profiles (Eddy, 1996) of all PKS domains and tries to put together a module based on rules for domain appearance in a module. It can also predict substrate specificity of AT domains thus enabling prediction of the synthesised polyketide chain. MAPSI (Tae *et al*., 2009) is a web application with a more complex structure than SEARCHPKS (Yadav *et al*., 2003) as its backbone consists of a database of all annotated polyketide clusters, which are used for identification of homologous clusters, as well as its "working part" storing information about currently analysed genomes. It also brings new functionality to PKS analysis software with a module for assembly of artificial polyketide synthases simply selecting modules with predefined domain composition and substrate specificity.

One of the recent additions to the field of biosynthetic gene cluster analysis programs is CLUSEAN – a computer-based framework for the automated analysis of bacterial secondary metabolite biosynthetic gene clusters (Weber *et al*., 2009). It integrates standard analysis tools like BLAST (Altschul *et al*., 1997) and HMMER (Eddy, 1998), with tools specific for the identification of the functional domains and motifs in NRPS/type I PKS and the prediction of specificities of NRPS. It is designed as a modular framework of BioPerl (Stajich *et al*., 2002) scripts which combine and manage results obtained from all the methods used. Combination of BLAST and HMMER is used to identify homologous proteins and to identify domains while the prediction of specificity of adenylation domains is based on approach developed for NRPSpredictor (Rausch *et al*., 2005)

1.2 Determining protein specificity/subgroups

Within one protein family several subgroups can exists which share the same interaction interface and mechanism but have different interaction partners. The ability to discriminate between different subgroups is gaining greater importance with the development of genetic/enzyme engineering, as can be seen on the example of PKS acyl transferase domains where changing the subgroup (defined by substrate specificity) of the domain results in synthesis of novel products. To be able to discriminate subgroups of a particular protein family, the group under consideration must be analyzed with respect to some relationship between its elements (usually homology), which then leads to a division of the group into a

number of disjoint subgroups. This approach is often referred to as clustering. With the advance of sequencing methods intensive studies on sequence clustering methods were also started. This primarily refers to various graph–based procedures, where one divides a group of sequences – coming from a single genome – into a certain number of disjoint groups of homologues (Donald and Shakhnovich, 2005; Abascal *et al*. 2002). Clustering can also be performed by the phylogenetic analysis, with branches of the tree corresponding to a hierarchical clustering scheme (Felsenstein, 2004).

Although clustering methods based on phylogenetic analysis are quite reliable in determining members of the subgroup, they lack the ability to determine amino acid residues determining this functional split. The other example in which phylogenetic analysis can fail is co-evolution of multiple features along with specificity in which other features can give a "stronger" phylogenetic signal than the desired specificity signal. With the increasing amount of genomic data a large number of methods which deal with this problem have been developed, of which only a few will be mentioned here. As a starting point they use the multiple alignment of the protein family in which they then, using various methods, search for residues specific for a subtype. It has to be pointed out that all of them require a set of predefined subgroups within the protein family in which then are able to recognise residues correlated with subgroup specificity. Hannenhalli and Russell (Hannenhalli and Russell, 2000) used relative entropy of the position (column of the multiple alignment) for each of the subtypes to estimate its role in determining the sub-type. For each position cumulative relative entropies for all subtypes were converted into Z-scores based on the distribution of entropies for an alignment. Z-scores are then used to assess each position's importance in determining the sub-types. Feenstra and co-workers (Feenstra *et al*., 2007) also used an entropy-based method, called *Sequence Harmony*, which is able to accurately detect subfamily specific positions from a multiple alignment by scoring compositional differences between previously defined subfamilies. Pazos and collaborators (Pazos *et al*., 2006) developed two methods for detecting positions which can incorporate external functional classification which may or may not coincide with the one implicit in the multiple sequence alignment (MSA). The *Xdet* method uses an external arbitrary functional classification instead of relying on the one implicit in the alignment to locate positions related to that classification. The *MCdet* method uses multivariate statistical analysis, based on vectorial

representation of sequences, residues and functions on related spaces, to locate positions responsible for each one of the functions within a multifunctional family. These methods are intended for protein families where a phylogeny/function disagreement is suspected. Wallace and Higgins (Wallace and Higgins, 2007) developed Between Group Analysis (BGA) – a supervised multivariate statistical method which can identify residues causing functionality change from families of proteins with different substrate specificities from respectable multiple alignment. The method is supervised in the sense that it requires sequences to be labelled as belonging to specific subgroups in advance. The BGA method is a graph method carried out in two steps. Firstly, the data set with defined different subgroups is ordered in vector space using either principal component analysis (PCA) or correspondence analysis (CA) so that similar objects are near each other and dissimilar further away. Secondly, the BGA finds linear combinations of the axes that maximise between-group variances and minimise within-group variances. The method has been successfully applied on a relatively small test set consisting of three protein families with results comparable to other methods but with an alternative method for viewing results (Wallace and Higgins, 2007). Kalinina and collaborators (Kalinina *et al*., 2004) introduced a method for automated selection of residues that determine the functional specificity of proteins with a common general function. Those residues are expected to be conserved within the orthologs (group assumed to have same specificity) and to vary between paralogs. The advantage of the method is taking directly into account nonuniformity of amino acid substitution frequencies and determining the thresholds automatically for each case. The SPEER method developed by Panchenko and co-workers uses several criteria to distinguish specificity-determining residues (Chakrabarti *et al*., 2007). It has a scoring function representing a linear combination of scores based on physico-chemical properties, evolution rate and combined relative entropy of amino acids. Based on a benchmark containing 13 protein families the method outperforms other methods tested, especially for marginally conserved sites and sites conserved in one subfamily and variable in another. One interesting observation in the paper was that from all the criteria used, the prediction accuracy mostly depends on the level of conservation of physico-chemical properties within the subfamily and between them.

A method capable of automatic protein subfamily identification and classification for the use in high-throughput applications has been recently developed by Brown and co-workers

(Brown *et al*., 2007). It is a pipeline which uses the Subfamily Classification in Phylogenomics (SCI-PHY) algorithm to automatically identify subfamilies and then to build a hidden Markov model (HMM) (Eddy, 1996) of the subfamily. SCI-PHY subfamilies closely correspond to functional subtypes defined by experts and to conserved clades found by phylogenetic analysis. Currently the PhyloFacts database (http://phylogenomics.berkeley.edu/phylofacts/) contains almost 60 000 family HMM profiles and more than 1,5 million subfamily-specific HMMs.

1.3 Computational structure prediction and protein-protein docking

As the function of a protein is defined by its structure, methods for determining protein structure from primary nucleotide/protein sequence are gaining greater importance since the next generation of annotation systems aims to integrate structural data into the annotation process (Reeves *et al*., 2009; Watson *et al*., 2005) as protein structure is much more conserved than the sequence during the evolution (Chothia and Lesk, 1986). The Structural Genomics Consortium started a project that aims to characterize the shapes and modes of action of the entire protein repertoire encoded within the genome. The plan is that with the advances in high-throughput X-ray crystallography and NMR methods to obtain a "dense set" of protein structures from which all other experimentally unsolved proteins would be in homology modelling range (http://www.thesgc.org/). Projects such as MODBASE (Pieper *et al*., 2009) are already carrying out this initiative through the combined use of PSI-BLAST (Altschul *et al*., 1997), for the identification and aligning of homologues and MODELLER (Eswar *et al*., 2007) for building of structural models.

Computational structure prediction methods can be classified into two general approaches. The first one includes threading and comparative (homology) modelling and relies on detectable similarity (usually minimally 30 % identity) of the modelled sequence with at least one known structure. The second type are *de novo* or *ab initio* methods which predict the structure from sequence alone, without relying on similarity at the fold level between the modelled sequence and any of the known structures (Baker and Sali, 2001). Homology modelling methods use the alignment of the query protein sequence (target) to one or more proteins with known structure (template) as a foundation for predicting the structure. The

entire homology modelling idea is based on the premise that sequences having certain sequence identity will also have similar structural folds. The structure prediction process consists of finding known structures related to the sequence to be modelled (template), aligning the sequence with the templates, building a model and evaluating the model (Marti-Renom *et al*., 2003). Templates for modelling are selected based on the sequence similarity with the target protein and the usual similarity search methods such as BLAST, PSI-BLAST (Altschul *et al*., 1990; Altschul *et al*., 1997) and profile HMMs (Eddy, 1998) are used.

For sequences that have templates with highly similar structure (minimum of 50 % sequence identity) fully automated structure prediction methods can be used (Arnold *et al*., 2006), while for the more difficult or unusual modelling cases better results are obtained with nonautomated, expert use of various modelling tools (Marti-Renom *et al*., 2003). For targets that do not have significantly similar templates fold recognition methods are used to predict structure. These methods are based on the principle that every distinct protein fold has its own pattern of features with which its amino acid sequence must be compatible and the target sequence is checked for compatibility with each of the various known folds. Mostly used are so-called threading methods where the 3-D to 1-D profile is created for each unique fold in the protein Data Bank (PDB). In each fold's 1-D profile every amino acid position in the protein is coded with a description of its environment based on physico-chemical properties (Bowie *et al*., 1991). The popular threading program GenTHREADER uses a variation of the 1-D profile with energy potentials instead of environment description for each position (Jones, 1999).

When there are no suitable templates the only possibility is to use *de novo* structure prediction methods which attempt to generate structural models solely on the basis of the principle of physics and chemistry. They start from the assumption that in the native state protein will be at the global free energy minimum and carry out a large scale search of conformational space for structures that are particularly low in free energy. Due to the vastness of conformational space that has to be searched most of the programs developed have had little success (Ginalski, 2006; Zhang, 2008). However, the Rosetta program package, with a slight variation of this approach, achieved very promising results. The method divides a protein into short segments (3 and 9 residues) which are continually

sampled in all possible local conformations until combinations with low energy, buried hydrophobic residues and paired β-strands are found. The search is greatly accelerated as free energy calculations are reduced because the fragments used to build the structure are taken from conformations found in experimentally determined structures so local interactions are close to optimal (Baker, 2006). The accuracy of *ab initio* approach fully depends on the method used while the accuracy of the comparative models is related to the percentage identity with the template. For targets with sequence identity higher than 50 % high-accuracy comparative models with about 1 Å root mean square (RMS) error for the main chain can be made. Medium-accuracy comparative models with 1.5 Å RMS error for the main chain can be obtained with targets having 30 -50 % identity to template. For targets having less than 30 % identity to template usually only low accuracy models can be obtained. These low-accuracy models besides mistakes in side-chains, core distortions and loop modelling errors, may also have entirely incorrect folds (Baker and Sali, 2001).

Protein-protein docking is the computational modelling of the quaternary structure of a protein complex starting from the individual structures of the individual proteins (Ritchie, 2008). Approaches used in protein-protein docking are trying to determine how proteins interact. From the analysis based on known yeast protein interactions it is estimated that each protein has roughly 9 interaction partners and that there are around 10,000 basic protein interaction types (Aloy and Russell, 2004). Although the number of experimentally determined protein structures has significantly increased in the recent years only a small number of those structures represent protein-protein complexes and it currently seems unlikely that it will be possible to apply high-throughput structural genomics techniques to protein complexes (Russell *et al*., 2004). That leaves computational techniques for the prediction of protein-protein docking as one method that might close the gap. Older methods, although still used and regularly achieving nearly-native docking solutions in tests (Janin *et al*., 2003), treated interacting proteins as rigid objects. As proteins are not rigid object these methods fail if there are large conformational changes in the docking process. To compensate for that limitation flexible protein-protein methods able to model both backbone and side-chain movements were introduced (Andrusier *et al*., 2008). In this work (Fig. 2) rigid body protein-protein docking methods (Comeau *et al*., 2004; Chen *et al*., 2003;

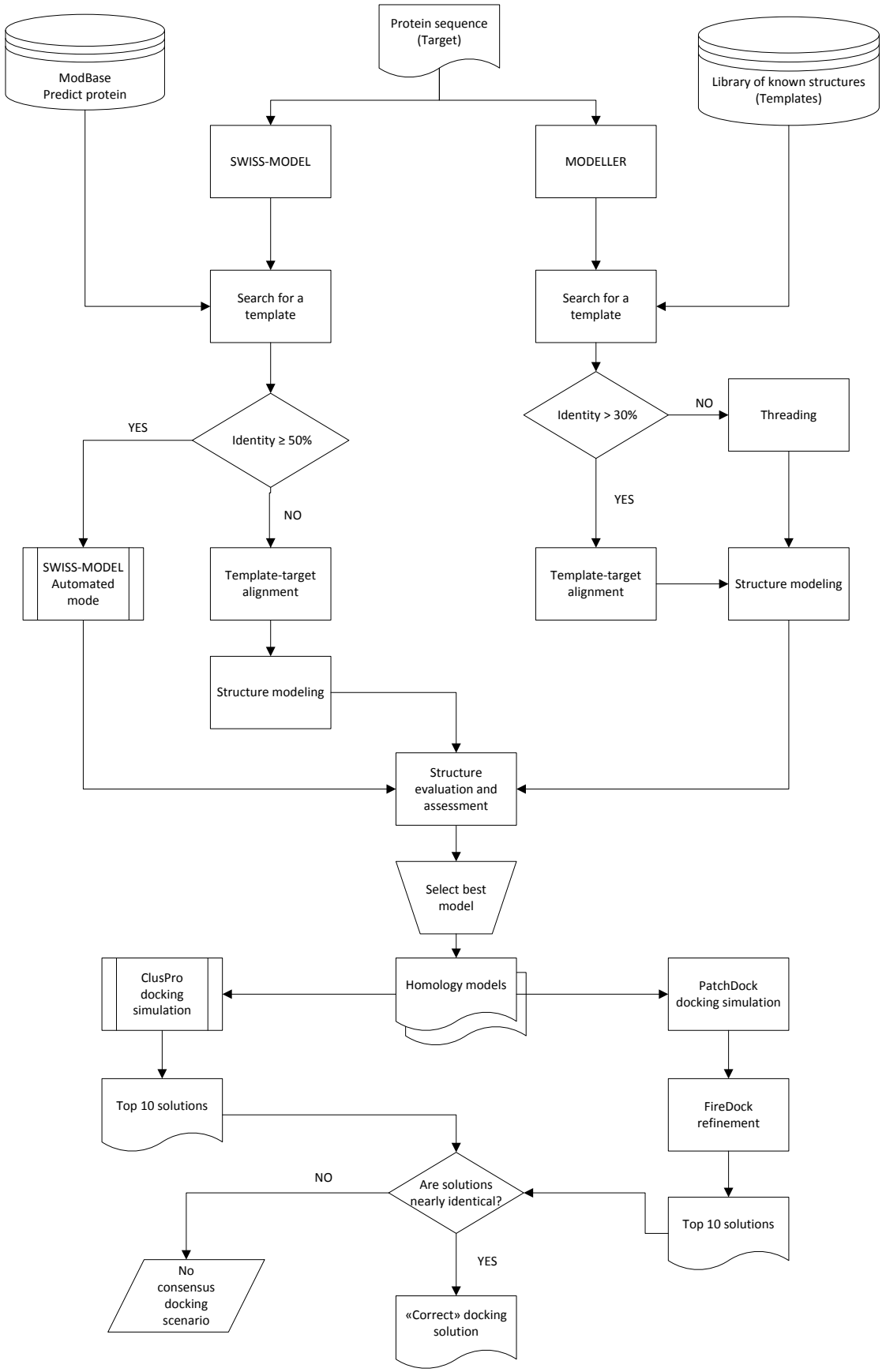

Fig 2 Flow chart representing homology modelling and protein-protein docking process of *in silico* interaction simulation of proteins constituting the daunorubicin polyketide synthase.

Schneidman-Duhovny *et al*., 2005) were used as it was hypothesized that during interaction proteins do not go through large conformational changes.

Some docking algorithms start by simplifying representation of the proteins by projecting each protein onto a regular 3D Cartesian grid, distinguishing grid cells based on their location (surface, core). Then a docking search is performed by scoring the degree of overlap between pairs of grids in different orientations. To speed up calculations by reducing the number of orientations to be searched a number of techniques are used, with Fast-Fourier Transform (FFT) correlation being the most popular. Besides grid-based FFT correlation algorithms the spherical polar Fourier (SPF) approach, that allows rotational correlations to be calculate rapidly, and the geometric hashing approach, in which the protein surface is scanned to create a list of critical points (pits, caps and belts) which are then compared in a clique-detection algorithm to create a small number of docking orientations for grid scoring, are also used (Ritchie, 2008). The methods mentioned above, generate a large number of docking solutions (from a few thousand up to tens of thousands) which should contain native (or near-native) solutions. Scoring functions implemented in these docking methods (desolvation, hydrophobicity, electrostatics) still have difficulties in distinguishing near-native solution within all the generated docking solutions. That is why some of the algorithms implemented a two step search and scoring procedure in which an initial list of docked solutions is re-scored using available biophysical information and information derived from analyses of existing protein-protein interfaces (Ritchie, 2008). The other approach that gave reasonable prediction, although not a scoring function, is the clustering of uniformly sampled low energy *ab initio* FFT docking solutions.

1.4 Annotation of genomic sequences

The process of identifying all genes that encode proteins for each genome as well as functional identification of as many proteins as possible is called genome annotation. Genome annotation is a two step process consisting of prediction of a gene's location in the genome based on characteristic sequence variations and identification of protein function by comparing it to already characterized proteins using various types of similarity searches. (Mount, 2004) These basic two steps of the annotation process can be carried out in various

modalities based on the direction of information flow, type of automatization, methods used to find genes and to infer their function and, most important, the organism which is being annotated. As the organisation of the prokaryotic and eukaryotic gene differs so differs the complexity of approaches used. Eukaryotic annotation systems have to deal with much more complex organisation of the gene as they have to predict correct intron-exon structure to be able to recreate coding sequences. For that purpose two programs have achieved best results - GeneWise and GeneScan. (Brent, 2005) GeneWise is the most important protein-to-genome alignment program whose accuracy mostly depends on the identity between protein and the locus to which is aligned to. On the protein identity range from 85 - 95 % to targeted locus it achieves more than 90 % exact exon specificity and 75 % exact exon sensitivity. (Birney *et al*., 2004) On the other hand, GenScan is *de novo* gene predictor that uses Generalized Hidden Markov model (GHMM) for predicting genes and exon-intron structure (Burge and Karlin, 1997). GenScan is currently, due to its accuracy, used as a genefinder in the Ensembl annotation pipeline (Curwen *et al*., 2004).

Concerning the flow of information in the annotation process the two approaches can be distinguished: **bottom-up** and **top-down**. More common is the bottom-up approach in which the individual base elements of the system are first specified in great detail and then linked together to form a larger subsystem which can be repeatedly linked until a complete top-level system is formed. In the bottom-up approach all know components and interactions to the model system are basically integrated. The top-down approach is essentially breaking down a system to gain insight into its compositional sub-systems. In the top-down approach an overview of the system is first formulated, specifying but not detailing any first-level subsystems. Each subsystem is then refined in ever greater detail, sometimes at many additional subsystem levels, until the entire specification is reduced to base elements. The critical difference in the two approaches occurs when all components and interactions are not known (Fraser and Marcotte, 2004). Based on the automatization level of the annotation process three types can be distinguished: manual, automatic and semiautomatic. First annotations were done manually by experts and were accurate but covered a relatively small percentage of genes in the genome and were slow. Some annotations are still done manually, mostly for reference databases such as VEGA (Wilming *et al*., 2008) and UNIPROTKB/SWISSPROT (The UniProt Consortium**,** 2010) which require combination of

detailed, accurate and critically assessed annotations inferred from genomic sequence combined with the information from journal publications and serve as a gold standard dataset for automated annotations systems relying on homology. With the advances in genome sequencing and nearing completion of Human Genome Project the first automated annotation systems started to appear. With time they incorporated more programs, methods and manually curated databases into their pipelines that were used to predict gene structure and function. Examples of those systems are the Ensembl annotation pipeline (Curwen, 2004), UCSC genome browser (Karolchik, *et al*., 2008) and the NCBI annotation pipeline (Wheeler, *et al*., 2007*).*

Most of the genes functions are inferred based on similarity with already annotated genes from manually curated databases. Various similarity search methods have been developed, the most notable one being BLAST (Altschul *et al.*, 1990) which has been in use for the last 20 years. In recent years more powerful methods such as PSI-BLAST (Altschul *et al.*, 1997) and Hidden Markov Models (HMMs) (Eddy, 1996) have been developed capable of identifying more remote protein homologues. One of the implementations of profile hidden Markov models (profile HMMs) for biological sequence analysis is a program suite called HMMER and it was used as the main similarity search method in this work. Profile HMMs are statistical models of multiple sequence alignments which capture position-specific information about conservation of each column of the multiple alignments (Eddy, 1998). Per column of the multiple alignment there is one match (M) state which emits a single residue with a probability score determined by the observed frequency of specific residue in the corresponding column of the multiple alignment. Each match state has an insertion (I) and deletion (D) state associated with it -the group of three states (M/D/I) at the same consensus position in the alignment is called a *node*. Each state has a state transition probability for transition to the next state. Transitions are arranged so that at each node either M state, which emits a residue, or D state, which does not emit a residue-resulting in gap character (-), can be used. Insertion (I) states occur between the nodes which can have, unlike M or D states, a self-transition which enables one or more inserted residues between consensus columns (Eddy, 2003).

The HMM architecture used in the HMMER programme package (version 2) – called Plan7 (Fig. 3), begins with dummy non-emitting begin state (B) and ends with dummy non-emitting end state (E). Between them is the core section of the model consisting of M, D and I nodes. This main model controls the data dependent features of the model while probability parameters are estimated from observed frequencies of residues and transitions in a multiple sequence alignment. The other states used in Plan7 (S, N, C, T, J) are control algorithm dependent features of the model (local alignment, multihit alignment etc.) and are set externally by the user rather than learned from the data (Eddy, 2003). HMMER is a program suite consisting of several programs from which *hmmbuild*, *hmmcalibrate*, *hmmsearch*, *hmmpfam* and *hmmalign* were used.

- *hmmbuild* creates a profile HMM from a multiple alignment of a protein (or nucleic acid) family. It also controls alignment style during the building of the profile HMM and supports the following modes: "glocal" (ls) - global with respect to the profile and local with respect to sequence, "multihit Smith-Waterman" - local with respect to both model and the sequence, "Smith-Waterman" classic local alignment - single best alignment per target, "Needleman-Wunsch" classic global alignment - single best hit per target.
- *hmmcalibrate* calibrates HMM search statistics by scoring a large number of synthesized random sequences to it, fits an extreme value distribution to the histogram of those scores and incorporates those data into the profile. Calibration of the profiles increases their sensitivity and makes detection of remote homologues more reliable.
- *hmmsearch* reads a profile HMM and searches sequence file for significantly similar sequence matches.
- *hmmpfam* reads a sequence file and compares each sequence in it against all the HMMs given to a program.
- *hmmalign* aligns a set of sequences from file to a HMM file and outputs a multiple sequence alignment.

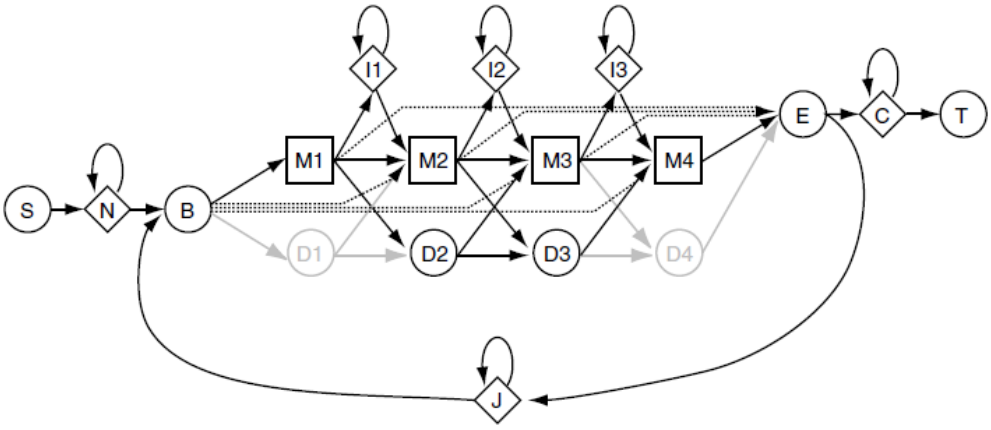

Fig. 3 The Plan 7 as implemented in HMMER. It consists of core section containing match (M), insert (I) and delete (D) states surrounded by two dummy non-emitting states - begin (B) and end (E). Core section and begin and end state make up the "main model" which controls the data dependent features of the model. Other states (S, N, J, C, T) are "special states" which control algorithm depended features of the model by parameters set externally by the user and not learned from the data. Image adopted from Eddy, 2003.

1.5 *Polyketide genes of Dictyostelium discoideum*

Dictyostelium discoideum is a species of social amoebae belonging to the eukaryotic phylum *Mycotozoa*, commonly called slime molds. Since *D. discoideum* has both single-cell and multicellular life stages it is used as a model organism for molecular mechanisms of cell motility, signal transduction, cell-type differentiation and developmental processes. In nature Dictyostelium inhabits forest soils where it feeds on bacteria and yeast which it tracks by chemotaxis (Eichinger et al., 2005). In times food abundance, *Dictyostelium* undergoes the vegetative cycle, preying upon bacteria in the soil and periodically dividing mitotically. With the depletion of food source, *Dictyostelium* enters the aggregation or social cycle. In the social cycle, amoebae aggregate by the thousands under the influence of a cAMP signal and form a motile slug, which moves towards chemoattractants such as light, heat and humidity. The cAMP and a polyketide derived differentiation-inducing factor (DIF) are involved in differentiation of the slug cells into prestalk and prespore cells. Ultimately the slug forms a fruiting body which consists of cellulose stalk and a spores-bearing sporangium (Fig. 4). Recently a third type of cells has been discovered – sentinel (S) cells with innate immune-like functions (Chen *et al*., 2007). Although mostly reproducing asexually, *D. discoideum* are under certain conditions (dark, humidity) capable of sexual reproduction. In

the sexual cycle, amoebae aggregate in response to cAMP and sex pheromones, and two cells of opposite mating types fuse, and then begin consuming the other attracted cells. Before they are consumed, some of the prey cells form a cellulose wall around the entire group, thus forming the giant diploid cell – macrocyst, which eventually undergoes meiosis and mitosis, and hatches hundreds of recombinants (http://dictybase.org/).

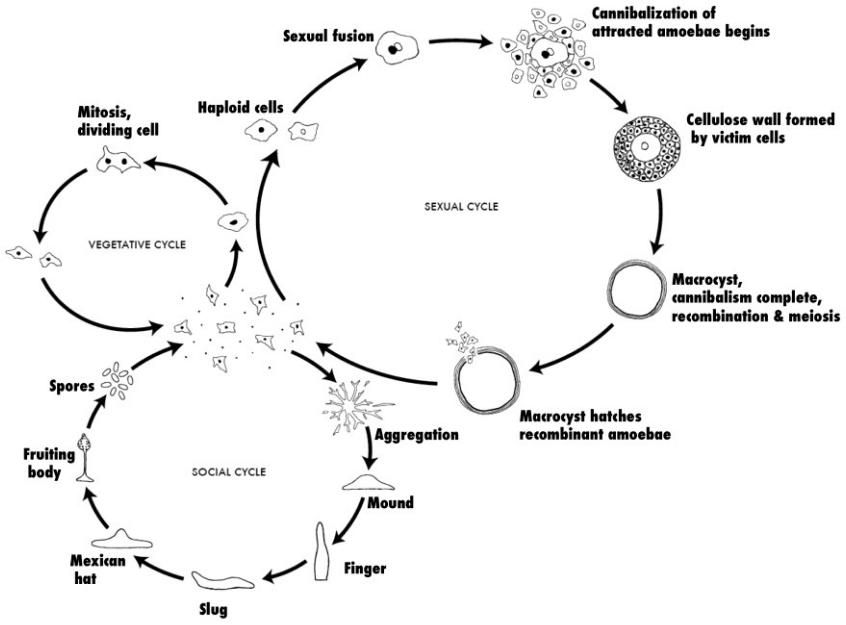

Fig. 4 Schematic presentation of the possible lifecycles of *Dictyostelium discoideum* (Source the dictyBase)

Except for its ability to alternate between unicellular and multicellular forms, *D. discoideum* is also noteworthy as representing one of the earliest branches from the last common ancestor of all eukaryotes. Although the taxonomical classification of *D. discoideum* has changed over time it is now believed that *Dictyostelium* diverged after the plant-animal split, but before the divergence of fungi (Bapteste *et al*., 2002). Despite the earlier divergence of *Dictyostelium*, many of its proteins are more similar to human orthologues than are those of *Saccharomyces cerevisiae* (Eichinger *et al*., 2005). Although plants, *Metazoa*, fungi and *Dictyostelium* all share 32 % of the eukaryotic Pfam domains, the majority of them are specific for certain group(s). Protein domains absent from plants and present in other groups are interesting as they probably arose after divergence of plants but before divergence of *Dictyostelium*. Most of the domains from this group of proteins are involved in cell cycle control and signalling (Eichinger *et al*., 2005).

The genome size of *D. discoideum* is 34 Mb and there are6 chromosomes. The genome is (A+T)-rich (77.57 %) and has a broadly uniform composition. On a finer scale there is a difference in A+T content of exons (73 %) and introns (88 %). The number of protein-coding genes in the genome is 13,605, from which 6,801 are represented by qualified ESTs (http://dictybase.org/). Interestingly in the *Dictyostelium* genome, unlike in human one, the number of genes was underestimated in the first analysis (8,000-10,000) and even at the time of publication of the genome sequence the number of genes was estimated as 12,500 (http://genome.imb-jena.de/dictyostelium/; Eichinger *et al*., 2005). The introns range in size from 66 to 2,298 nucleotides with an average size of 146 nucleotides, and occur in more than two thirds of genes. The 5' splice site has a /GTAAGT consensus sequence, with the dinucleotide/GT being completely conserved. The 3' splice site has a ATAG/ consensus sequence, again with the dinucleotide next to the splice site (AG/) completely conserved. The branching site is not very conserved in *Dictyostelium* as only half of the introns had the conserved CTNA consensus element found in yeast and higher eukaryotes (Rivero, 2002).

From the whole genome annotation a few gene families were found in great numbers, namely PKS and ABC transporters. Until now *Dictyostelium* is the organism with the highest number of PKS genes (Ghosh, 2008). However, the functional relevance of most of these genes remains quite obscure. Until now only two of the PKS genes have been functionally characterized (Steely1 and Steely2 - pks1 and pks37 in the dictyBase naming scheme respectively). Both of them are proposed to produce differentiation-inducing factors (DIFs) - compounds involved in orchestration of cell differentiation. DIFs and DIF analogs have recently attracted medical interest as it was shown they are able of inhibiting cell proliferation and induce differentiation of mammalian cells (Gokan *et al*., 2005). What initially attracted attention to these genes is their unusual organization – they resemble type I PKS genes, but, instead of a final thioeseterase domain (TE), they both end with a type III PKS (chalcon synthase). This is the first time that such PKS I-PKS III hybrid has been found. (Austin *et al*., 2006) Although the exact biosynthetic pathway is still not unambiguously elucidated it is believed that type I PKS part of the protein uses acetyl-CoA as a starter unit and malonyl-CoA as extender units to synthesize acyl-thioester intermediaries which are then passed to the type III PKS. Acyl intermediaries are used by the type III PKS as a starter unit upon which several elongation cycles (pks1 two and pks37 three elongation cycles) are

carried out. (Ghosh *et al*., 2008) All other Dictyostelium PKS genes structurally and organisationally resemble to the iterative type I PKS and most likely follow the biosynthetic logic of iterative type I PKS.

1.6 Goals and work objectives

To predict the compound synthesised by PKS genes is, until now, possible only for modular type I PKSs as, for the other classes, mechanisms controlling biosynthesis are still not fully understood. Modular type I PKSs function as assembly lines and by dissecting all of the components involved in that machinery one can predict the aglycon synthesized. Our group has been involved in development of a software package which was intended to predict the chemical molecule synthesized by the enzyme based solely on genetic information coding for it. To do that PKS gene(s) have to be firstly identified, all domains present have to be determined together with their respective activity and specificity (if one exists) and then all this data has to be assembled to allow prediction of all the building units incorporated into the polyketide chain, their respective reduction level of the keto group and finally the structure of the synthesized polyketide itself.

In some cases a domain may have several potential substrates, for example, different PKS acyltransferase (AT) domains can incorporate different building units into the polyketide chain. To be able to distinguish the specificity of a domain several methods were already available. They split a protein family into subgroups, but do not define the residues causing this change of function, or are able to detect specificity determining residues (SDR) but only on a protein family already divided into subgroups. The aim was to develop a method to cluster the family into functional groups without prior knowledge.

The mechanisms controlling biosynthesis of type II PKSs are not fully understood and it is not possible to predict the synthesized product. It is believed that interaction of the subunits of the PKS complex plays an important role but until now the only complex structure which has been solved is that of two ketosynthase subunits (KS and CLF). Using protein-protein docking simulations interactions between subunits of the complex will be investigated aiming to give information about all interacting partners and the complexes formed.

As the search for novel natural products remains one of the cornerstones of the quest for novel active compounds, new ecological niches and organisms more distant from the ones currently used as producers of natural products are being examined. *Dictyostelium* was the first of the slime moulds to have a genome sequence and it showed a genome especially rich in PKS genes. In comparison with the most popular producers of active compounds used by the pharmaceutical industry, the genus *Streptomyces*, it contained twice the number of PKS genes none of which has been assessed for medicinal purposes. The analysis of the PKS genes also showed errors done in automatic annotation of its genome due to the presence of introns. The other interesting characteristics of *Dictyostelium* were its position in the tree of life (phylogenetic tree), which showed certain incongruities based on the type of analysis, and its lifecycle, which shows high complexity for a unicellular life form, certainly giving hope that *Dictyostelium* might contain unique features not commonly present in other life forms. A more detailed analysis of the genes might reveal more information about their origins, functions and possible usefulness as producers of novel natural products.

Scientific papers

2. Scientific papers

2.1 *Clustering of protein domains for functional and evolutionary studies*

Pavle Goldstein, **Jurica Zucko**, Dušica Vujaklija, Anita Krisko, Daslav Hranueli, Paul F. Long, Catherine Etchebest, Bojan Basrak and John Cullum. *BMC Bioinformatics*, ***10***, 335, 2009.

Abstract:

Background

The number of protein family members defined by DNA sequencing is usually much larger than those characterised experimentally. This paper describes a method to divide protein families into subtypes purely on sequence criteria. Comparison with experimental data allows an independent test of the quality of the clustering.

Results

An evolutionary split statistic is calculated for each column in a protein multiple sequence alignment; the statistic has a larger value when a column is better described by an evolutionary model that assumes clustering around two or more amino acids rather than a single amino acid. The user selects columns (typically the top ranked columns) to construct a motif. The motif is used to divide the family into subtypes using a stochastic optimization procedure related to the deterministic annealing EM algorithm (DAEM), which yields a specificity score showing how well each family member is assigned to a subtype. The clustering obtained is not strongly dependent on the number of amino acids chosen for the motif. The robustness of this method was demonstrated using six well characterized protein families: nucleotidyl cyclase, protein kinase, dehydrogenase, two polyketide synthase domains and small heat shock proteins. Phylogenetic trees did not allow accurate clustering for three of the six families.

Conclusions

The method clustered the families into functional subtypes with an accuracy of 90 to 100 %. False assignments usually had a low specificity score.

Own contribution to the paper:

Analysis of predicted specificity-determining residues and clustering for acyltransferase, dehydrogenase, ketoreductase, nucleotidyl cyclases and protein kinase protein families. Analysis of clustering stability based on motif length and comparison of clustering results with results obtained using phylogenetic methods.

BMC Bioinformatics

BioMed Central

Research article

Open Access

Clustering of protein domains for functional and evolutionary studies

Pavle Goldstein[1], Jurica Zucko[2,5], Dušica Vujaklija[3], Anita Kriško[4,7], Daslav Hranueli[5], Paul F Long[6], Catherine Etchebest[8], Bojan Basrak[1] and John Cullum*[2]

Address: [1]Department of Mathematics, University of Zagreb, Bijenicka 30, 10000 Zagreb, Croatia, [2]Department of Genetics, University of Kaiserslautern, Postfach 3049, 67653 Kaiserslautern, Germany, [3]Department of Molecular Biology, Rudjer Boskovic Institute, Bijenicka 54, 10000 Zagreb, Croatia, [4]Mediterranean Institute for Life Sciences, Mestrovicevo setaliste bb, 21000 Split, Croatia, [5]Faculty of Food Technology and Biotechnology, University of Zagreb, Pierottijeva 6, 10000 Zagreb, Croatia, [6]The School of Pharmacy, University of London, 29/39 Brunswick Square, London, WC1N 1AX, UK, [7]INSERM U- 571, Faculté de Médecine, Université Paris V, 156 rue de Vaugirard, 75730 Paris Cedex 15, France and [8]Equipe de Bioinformatique Génomique et Moléculaire, INSERM U-726, Université Denis Diderot - Paris 7, 2 place Jussieu, 75251 Paris Cedex 05, France

Email: Pavle Goldstein - payo@math.hr; Jurica Zucko - jzucko@pbf.hr; Dušica Vujaklija - vujaklij@irb.hr; Anita Kriško - akrisko@irb.hr; Daslav Hranueli - dhranueli@pbf.hr; Paul F Long - paul.long@pharmacy.ac.uk; Catherine Etchebest - catherine.etchebest@univ-paris-diderot.fr; Bojan Basrak - bbasrak@math.hr; John Cullum* - cullum@rhrk.uni-kl.de

* Corresponding author

Published: 15 October 2009

BMC Bioinformatics 2009, **10**:335 doi:10.1186/1471-2105-10-335

Received: 14 February 2009
Accepted: 15 October 2009

This article is available from: http://www.biomedcentral.com/1471-2105/10/335

Abstract

Background: The number of protein family members defined by DNA sequencing is usually much larger than those characterised experimentally. This paper describes a method to divide protein families into subtypes purely on sequence criteria. Comparison with experimental data allows an independent test of the quality of the clustering.

Results: An evolutionary split statistic is calculated for each column in a protein multiple sequence alignment; the statistic has a larger value when a column is better described by an evolutionary model that assumes clustering around two or more amino acids rather than a single amino acid. The user selects columns (typically the top ranked columns) to construct a motif. The motif is used to divide the family into subtypes using a stochastic optimization procedure related to the deterministic annealing EM algorithm (DAEM), which yields a specificity score showing how well each family member is assigned to a subtype. The clustering obtained is not strongly dependent on the number of amino acids chosen for the motif. The robustness of this method was demonstrated using six well characterized protein families: nucleotidyl cyclase, protein kinase, dehydrogenase, two polyketide synthase domains and small heat shock proteins. Phylogenetic trees did not allow accurate clustering for three of the six families.

Conclusion: The method clustered the families into functional subtypes with an accuracy of 90 to 100%. False assignments usually had a low specificity score.

BMC Bioinformatics 2009, **10**:335 http://www.biomedcentral.com/1471-2105/10/335

Background

Rapid progress in DNA sequencing is generating large numbers of deduced protein sequences. The prediction of their function is an important problem in Bioinformatics. This is tackled by comparing new sequences to known sequences as high sequence similarity usually indicates related function. It is possible to use similarity search algorithms such as BLAST [1]. A more sensitive approach is to use hidden Markov models (HMMs) to define protein families as implemented in HMMER suite of programs [2]. Such HMM profiles are used to define protein families in the Pfam database [3]. In many cases, these families consist of functional domains in larger proteins.

In many cases protein families can be split into sub-types based on functional differences e.g. substrate specificity such as for the malonyl-CoA- and methylmalonyl-CoA-incorporating acyl transferase domains of modular polyketide synthetases [4,5]. These differences usually correlate with specific differences in amino acid sequence, which help to understand the molecular basis of protein function and serve as a basis for building prediction programs [6]. In order to identify such diagnostic amino acids, it is first necessary to produce a multiple alignment of the protein sequences to identify corresponding residues in different members of the family. This can be done in various ways e.g. using an HMM-profile [2] or a multiple alignment program such as ClustalW [7]. In some cases, it is possible to identify diagnostic residues merely by inspection of sequences (e.g. [8,9]), but this is difficult or impossible in many cases.

An interesting approach that analysed the entropy associated with different residue positions was described by Hannenhalli and Russell [10]. The biological idea behind this approach is that amino acid residues that are important in the determination of functional subtypes will have different constraints depending on the subtype. In general they will not be absolutely conserved, but evolution will only allow limited variation and the pattern of variation will be different for different subtypes. The functional subtypes corresponding to each protein are input to the program and the program uses an entropy measure to identify residues that split the dataset between the functional subtypes. The detection of specificity-determining residues has been developed further [11-14]. The residues identified by these methods can be used to assign new sequences to the correct subtype. However, it must be emphasized that all these methods rely on experimental data about the subtypes of a sufficiently large collection of proteins to identify the residues.

In many cases of interest there may not be enough experimental data about subtypes, but there is usually a much larger set of protein sequences (deduced from DNA sequences) which have not been experimentally characterised. In this paper we describe a method which divides a set of protein sequences into subtypes based solely on sequence data without any prior assignment of subtypes. The method clustered six well-characterised protein families into functional subtypes without any prior knowledge of protein properties and identified specificity-determining amino acid residues.

Results and Discussion

Identification of subtypes

The starting point for the analysis was a multiple sequence alignment of the protein family being analysed. We used ClustalW and ClustalX [7,15] to align sequences [see Additional file 1]. Any other method of generating multiple alignments could be used e.g. with an HMM-profile of the family as implemented in the HMMER suite of programs [2]. The program only considers columns in the multiple alignment which contain amino acids for every member of the protein family (i.e. positions with any gaps are ignored). The program analyses the amino acids present at a given position and performs a statistical test to determine whether the distribution of the amino acids is more compatible with a model that they cluster around a single amino acid or with a model that they cluster around two or more different amino acids; the number of clusters is given to the program as a parameter. The two amino acid model has proved most useful for the six cases considered in this paper i.e. a binary split of the family into two subtypes is attempted. The statistical test needs a model for the substitution of amino acid residues and the BLOSUM-50 matrix [16] was used, which represents the observed substitutions in a large sample of proteins. Although this model will not be strictly true for each amino acid position, the success of the program shows that it is adequate. An evolutionary split statistic was defined (see Methods) that measures how well the position fits the multiple amino acid model i.e. a large value of the statistic indicates that the position should be important in the discrimination between subtypes.

On the basis of the evolutionary split statistic, the user selects a series of positions (a "motif") to be used for splitting the protein family into subtypes. These are typically positions with the best scores, but other criteria (e.g. residues in a particular region or residues close to the active site if a 3-D structure of a family member or a related protein is available) can be used. The clustering algorithm used gives log likelihood values for each sequence that show how well the "motif" assigns the sequence to a particular class. When a division into two subtypes is being carried out, it is useful to use the "specificity score", which is the difference between the log likelihoods for assignment to the two classes. The specificity score is a measure of how good the assignment to the class with higher like-

lihood is. The user can experiment with different numbers of motif positions to find a selection that gives good discrimination. As we will show later, in most cases this choice is not critical for the success of the method.

Performance of the program

The program was tested on six different protein families [see Additional file 2]. Nucleotidyl cyclases have two functional subtypes corresponding to use of the substrates ATP or GTP respectively. We extracted 75 sequences (33 adenylate cyclases, 42 guanylate cyclases) from the UniProt database [17]. When the five positions with the best evolutionary split statistic were used to divide the family into two subtypes, the resulting groups were exactly the adenylate and guanylate cyclases (100% accuracy). Five of the ten best positions corresponded to amino acids that were discussed by Hannenhalli and Russell [10] as important in determining the functional subtype (Table 1).

The protein kinase family can be divided into serine/threonine and tyrosine kinases. 215 kinase sequences (85 serine/threonine, 130 tyrosine) were extracted from the protein kinase resource database [18]. When the 7 best positions were used, the program divided the kinases into subtypes with 100% accuracy. Seven of the best ten positions were identified previously as important for the subtype determination [10].

Lactate (LDH) and malate (MDH) are subtypes of a large dehydrogenase family. They show considerable sequence variability [19] making them a more difficult case than the first two families. 183 dehydrogenase sequences (74 LDH and 109 MDH) were extracted from the UniProt database [17]. When the top 6 positions were used as a motif the dehydrogenases were split into an LDH and an MDH group with 5 wrong assignments (97% accuracy). The wrong assignments all had low specificity scores (Figure 1).

The two residues with the highest evolutionary split scores were discussed by Hannenhalli and Russell [10] as important in determining the functional subtype. Experimentally it has been shown that a major determinant of the substrate specificity is the choice between glutamine or arginine at residue 144 (residue 102 of [19]). This position was the 14th best evolutionary split score in our analysis (Figure 2). The reason why it does not rank higher is that arginine/glutamine exchanges are fairly common in proteins (and have a score of +1 in the BLOSUM-50 matrix used by the program).

The acyl transferase (AT) domains of Type I modular polyketide synthases (PKS) determine the substrate selection [4,5,20-23]. Most incorporate either a C2 unit (malonyl-CoA substrate) or a C3 unit (methylmalonyl-CoA

Table 1: Nucleotidyl cyclases: residues with best evolutionary split scores.

	Residue number in multiple alignment		Substrate	
Evolutionary split score	**This paper**	**Hannenhalli and Russell, 2000**	**ATP**	**GTP**
113	1509	-	**C**	**V**
110	1636	1020	**W**	**F**
110	1634	1018	**D**	**C**
109	1630	1014	**K**	**M**
91	1517	919	**I**	**Y**
86	1580	-	**F**	**M**
84	1533	935	**E**	**Y**
83	1440	-	**M**	**E**
83	1497	-	**C**	**Y**
81	1656	-	**H**	**Q**

The ten residues with the best evolutionary split scores in the multiple sequence alignment of the nucleotidyl cyclases. When the residue had been detected in previous work [10] the corresponding residue number is given. The dominant amino acid for the two subtypes is shown.

BMC Bioinformatics 2009, **10**:335 http://www.biomedcentral.com/1471-2105/10/335

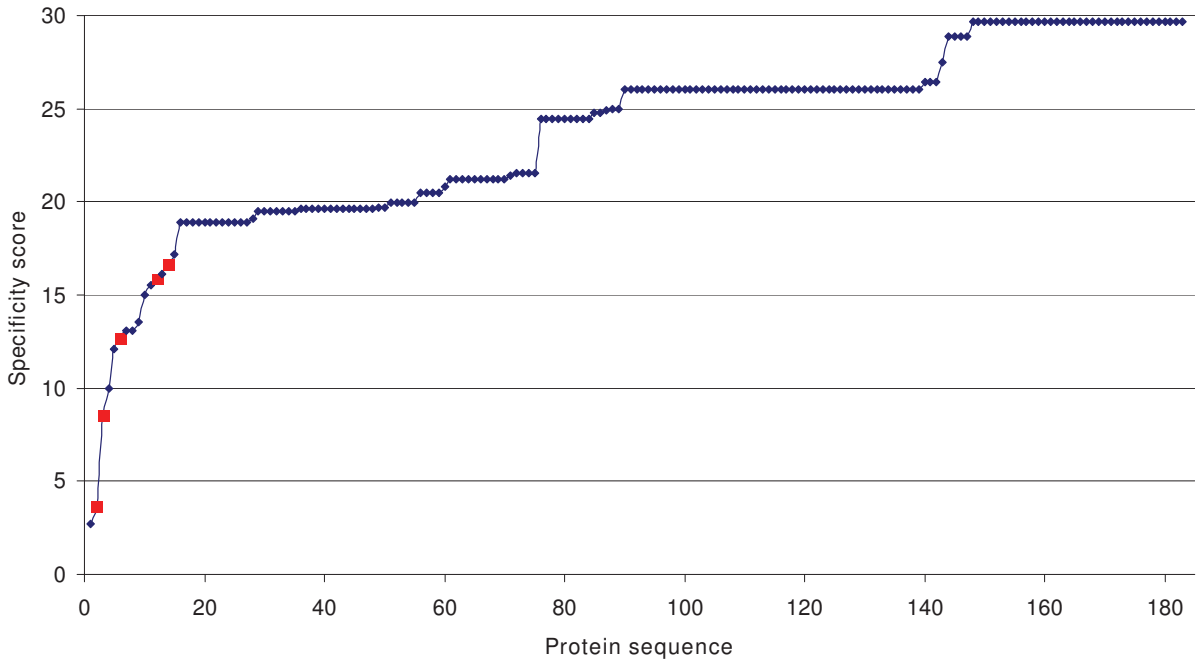

Figure 1
Specificity scores for the dehydrogenase family. The 183 LDH and MDH sequences are ordered according to specificity scores. The five wrongly assigned sequences are indicated in red.

substrate). The choice of substrate can be deduced from the chemical structure of the polyketide product. We chose 177 AT domains (99 C2, 78 C3). We used the top 7 positions to define a motif and the program divided the domains into C2 and C3 subtypes with only 5 wrong assignments (97% accuracy). The wrong assignments all had low specificity scores (among the lowest 6 scores of the 177 sequences). The top 7 amino acid positions chosen were positions previously recognized by Yadav and collaborators [9] by inspection of the sequences. The top 30 amino acid residues were identified in the sequence of *Escherichia coli* fatty acid synthase AT for which a 3-D protein structure has been determined ([24]; PDB ID 1MLA). The top 7 residues are in the region of the binding pocket where a direct effect on substrate binding might be expected. The other residues are scattered on the surface of the protein, too far from the substrate binding pocket to have a direct effect.

Ketoreductase domains (KR) of Type I modular PKSs use NADPH to stereospecifically reduce the initially formed keto group to a hydroxyl group [25]. The stereospecificity can only be deduced from the structure of the product for cases in which further reduction steps have not occurred. We used 72 KR domains for which the stereospecificity was known (33 R and 39 S). In this case, most of the residues with the best values for the evolutionary split parameter were clustered in a region of the sequence, so we chose the residues from positions 114 to 155 of the alignment to split the family into subtypes. This gave a 90% accurate assignment of domains (7 domains were misclassified). The motif residues included the residues that had been recognized by Caffrey [8] as playing a role in stereospecificity.

The final family that we examined was the small heat shock proteins (sHSP), where it is not clear whether there is a functional difference between different subtypes. We analysed 214 sequences and on the basis of the best four positions obtained a split between metazoan sHSPs and the others (plants, fungal, eubacterial and archaebacterial) (95% assignment) which corresponds to previously reported phylogenetic results [26]. The four residues (alignment positions 274, 292, 406 and 408) were localized on the 3-D structures of sHSPs from *Triticum aestivum* [27] and *Methanococcus janaschii* [28]. The four residues are in a region of the protein that is involved in dimerisation. It is known that oligomerisation is important for the function of the protein and this result suggests that the two subtypes identified might differ in oligomerisation properties.

The clustering algorithm allows a free choice of amino acids alignment positions to include in the motif. This

BMC Bioinformatics 2009, **10**:335 http://www.biomedcentral.com/1471-2105/10/335

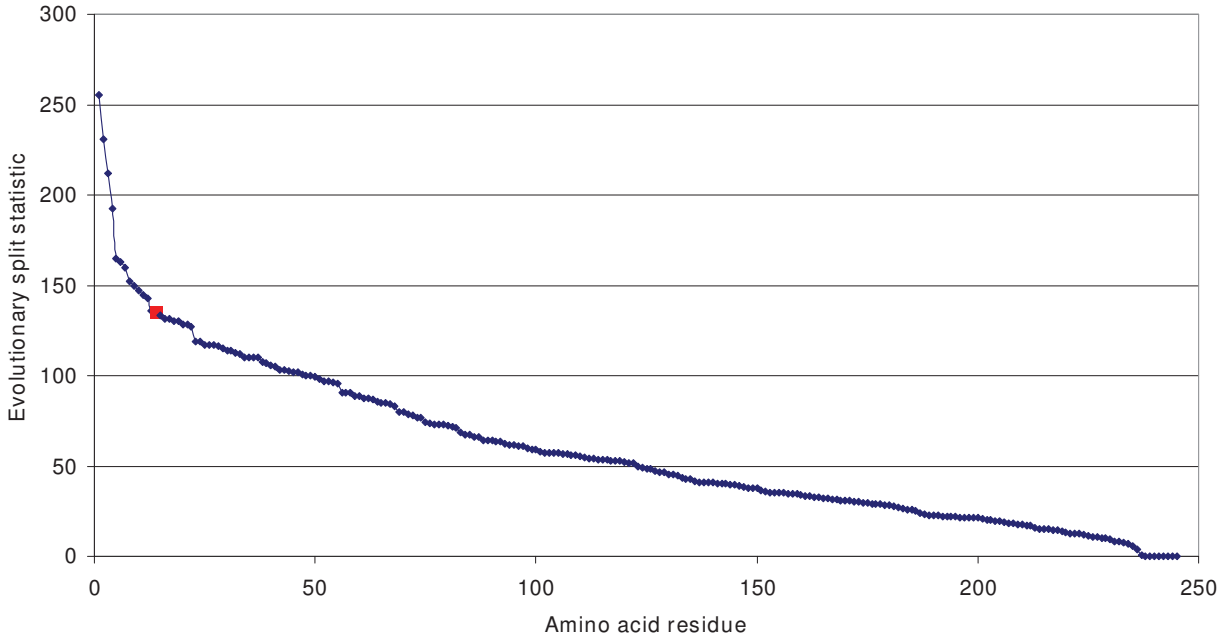

Figure 2
Evolutionary split scores for amino acid residues of the dehydrogenase family. The amino acid residues in the LDH/MDH multiple alignment are ordered using the evolutionary split score. Residue 144 of the alignment (Q in LDH, R in MDH) is shown in red.

raises the question as to how sensitive the clustering algorithm is to the exact choice of motif. We clustered the six protein families using amino acid positions with the highest evolutionary split scores and varying the length of the motifs from 5 to 30 positions. Table 2 shows that the accuracy of the clustering does not depend strongly on the number of positions chosen. This means that the algorithm could be used for the automatic clustering of protein families using a standard length of motif chosen from the best evolutionary split scores. In the case of the KR-domains, choosing a segment of the protein on the basis of specific knowledge, as done above, gave better results than using the best evolutionary split scores. The assignment of KR domains to subfamilies is complicated as they also determine the stereochemistry of methyl groups [29] and examination of 3-D structures of KR domains resulted in their division into six subtypes [30].

Table 2: Effect of motif length on clustering performance.

	motif length						
	5	**10**	**15**	**20**	**25**	**30**	
Protein family	**false assignments**						**No. sequences**
Nucleotidyl cyclases	0	0	0	0	0	0	75
Protein kinases	0	0	0	0	0	0	215
MDH/LDH	5	6	5	5	4	4	183
AT-domains	2	3	4	4	5	5	181
KR-domains	20	18	20	17	10	9	72
sHSP	10	13	14	11	5	5	214

The amino acids positions with the highest evolutionary split scores were used to construct the motifs.

The evolutionary split statistic allows the identification of residues that are important for the determination of subtypes. However, as it is calculated independently of the clustering, it is not as good as methods that are based on a known clustering. The subfamilies predicted by our clustering algorithm can be used for such analyses [10-14], which will give a more accurate identification of residues important for division into subtypes. The omission of sequences with low specificity scores should improve the analyses by removing misclassified sequences.

The algorithm showed an efficient division into subtypes for the six protein families tested. An alternative approach

to recognizing subtypes in the absence of functional information is to use phylogenetic analysis. In order to have a closer comparison with our clustering algorithm we constructed phylogenetic trees from the multiple alignments of our six protein families using distances calculated from a BLOSUM matrix [31] instead of the more common JTT method [32]. For the nucleotidyl cyclases and protein kinases, whose subtypes were recognised with complete accuracy by our method (Table 2), the functional subtypes do form separated clusters in the phylogenetic trees [Figure 3(A)) and 3(B)]. Division of the sequences into two subfamilies implies choosing a rooting point in the tree so that the subfamilies become clades in the rooted tree. In neither case, is the choice of such a rooting point unambiguous. For the cyclases [Figure 3(A))] there are several plausible rooting points, only one of which will give the correct subfamilies. The kinases [Figure 3(B))] fall into three clusters and the phylogenetic tree does not suggest the correct split into the two functional subtypes. The dehydrogenases [Figure 3(C))] also appear to split into three clusters and the phylogenetic tree does not suggest a division corresponding to the two functional subtypes, whereas our clustering program recognises the functional subtypes efficiently (Table 2). The AT-domains [Figure 3(D))] can be recognised as two groups using the phylogenetic tree with a similar degree of error to the clustering algorithm. The subtypes of the KR-domains [Figure 3(E))] cannot be recognised using the phylogenetic tree, whereas the two subtypes of the sHSPs are clear in the phylogenetic tree [Figure 3(F))]. Thus, in three of the six families, the phylogenetic trees did not give a clear identification of the functional subtypes. A further major advantage of the clustering algorithm is that the specificity score identifies sequences that are not well clustered by the algorithm so that they can be removed or treated with caution in subsequent analyses. The tests with known families showed that most wrong assignments involved such sequences.

In principle, the programs can also be used to cluster sequences into three or more subtypes. We tested for a clustering into three subtypes using two protein families: 92 serine proteases (67 trypsin-, 17 chymotrypsin-, 8 pancreatic elastase-subfamilies) and 59 AT-domains (28 incorporating methylmalonate, 18 malonate and 13 methoxymalonate). Clustering was undertaken using best 10, 20 and 30 positions for the evolutionary split statistic (data not shown). The clustering did not show a strong dependence on the number of positions. For the serine proteases, the trypsin subfamily was split into two groups and the chymotrypsin and pancreatic elastase subfamilies clustered together giving wrong clustering of 42 of the 92 sequences. Similarly, 22 of the 59 AT-domains were wrongly clustered. Thus, although the method works for carefully constructed sets of test data, it does not seem to be effective for real biological protein families. It is not surprising that the method becomes less effective with increasing number of subtypes. The potential of a column to contribute towards a *k*-way split is estimated with the evolutionary split statistic (formula 7) and increasing the number of subtypes drastically increases dimensionality of the parameter space; i.e. it is increasingly difficult to distinguish between evolutionary noise and functionally significant mutations. Thus, only exceedingly large sample sizes will provide sufficient power for the method to work well. Clustering is most efficient when the different subtypes are present in comparable numbers and the examples analysed in this paper show that the known sequences in natural protein families can often fall into one or two major subtypes with other subtypes being rare. Such situations can be analysed better by using binary clustering and subsequently looking for rarer subtypes in the sequences that have low specificity scores.

The method suffers from the drawback that it can only be used in practice for dividing protein families into two subtypes. This will cause problems for protein families with several common subtypes and the method may not work well for rare subtypes. Now that the feasibility of such a clustering algorithm has been demonstrated it is likely that improved algorithms can be devised to overcome these problems.

An important practical advantage of our algorithm is that it is computationally efficient allowing implementation on a public server. Using a standard PC with a 2 GHz processor, it needs about 0.1 second per column to compute the evolutionary split parameter (nearly independently of the number of sequences) and about 1 minute to compute the clustering into subtypes. It is therefore feasible to experiment with different motifs and different selections of the sequences to obtain optimal results. The method offers a useful tool to detect previously unsuspected clustering into subtypes. If experimental data for a limited number of proteins are available, they provide an independent test for the predicted clustering and the subtype of previously uncharacterised proteins is predicted.

Conclusion

The programs cluster protein families into subtypes effectively without any prior functional knowledge. The specificity score identifies protein sequences that do not cluster well into the defined subtypes: these may include further rare subtypes. The programs are especially suitable for detecting novel unsuspected subtypes where extensive sequence data, but little experimental data are available.

Methods

Preparation of sequences

The amino acid sequences for 75 nucleotidyl cyclases, 183 dehydrogenases and 214 small heat shock proteins

BMC Bioinformatics 2009, **10**:335 http://www.biomedcentral.com/1471-2105/10/335

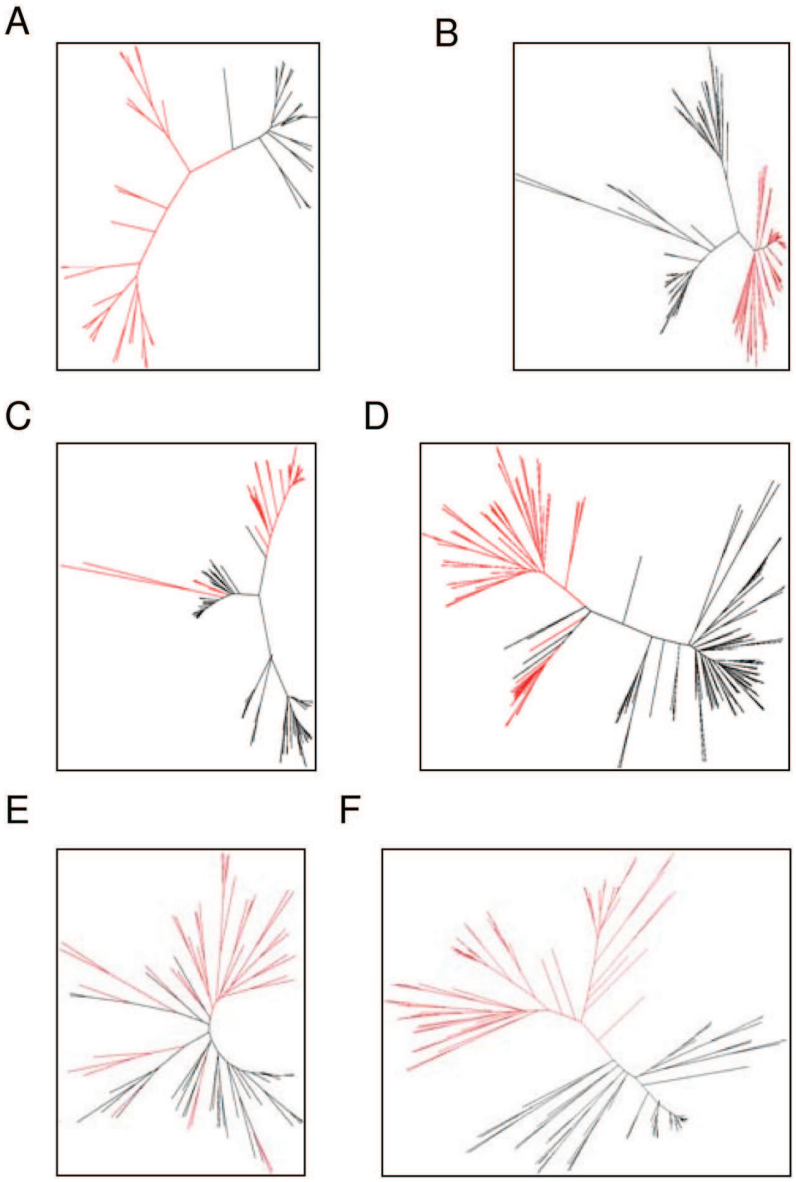

Figure 3
Phylogenetic trees of the protein families. The alignments of six protein families were used to construct phylogenetic trees from distances based on a BLOSUM matrix using a minimum evolution criterion. In each case, the branches corresponding to one of the two subfamilies are coloured red. (A) nucleotidyl cyclases (guanylate red), (B) protein kinases (tyrosine red), (C) dehydrogenases (LDH red), (D) AT-domains (C3 red), (E) KR-domains (S stereochemistry red), (F) sHSPs (metazooan black, others red).

(sHSP) and 92 serine proteases were extracted from the UniProt database release 53.0 or 57.0 [33]. The amino acid sequences for 177 acyltransferase (AT) and 72 ketoreduction (KR) domains from modular polyketide synthases were obtained from the NRPS-PKS database [34,35]. The amino acid sequences of 85 serine/threonine and 130 tyrosine protein kinases were retrieved from the protein kinase database [18]. All 59 AT-domains were extracted from the following clusters: ascomycin, concanamycin, FK506, geldanamycin, herbimycin, niddamycin, soraphen using the MAPSI database [36]. Multiple alignments of the sequences were constructed using ClustalW and Clustal X [7,15,37]. These multiple alignments for each family are shown in additional materials.

Construction of phylogenetic trees

Phylogenetic trees were constructed from the multiple alignments using the neighbour joining algorithm in version 3.66 of the PHYLIP package [38]. The distances were calculated with the Protdist program using the PMB (Probability Matrix from Blocks) model [31].

A model of amino acid substitutions

Let A be the alphabet consisting of twenty standard amino acids, and let $q = (q_1, ..., q_{20})$ be the stationary (marginal) distribution of elements of A in some protein universe P. We denote by $e_i = (0,...,1,...0)$ the i-th vector in the canonical basis of R^n, with 1 at the i -th position, and zeros elsewhere.

Definition 1 A *substitution model* for P is a family of distributions $a_{i,t}$, $i \in A$, $t \in [0, \infty\rangle$, that, for each $i \in A$, satisfies

$$\lim_{s \to 0} a_{i,s} = e_i$$

$$\lim_{t \to \infty} a_{i,t} = q.$$

Here $q = (q_j)$ is the vector of frequencies with which amino acids occur in the family of proteins. For $a_{i,t} = (a_{i,t}(1), ... a_{i,t}(20))$, $a_{i,t}(j)$ is, by definition, the probability of amino acid i mutating into j after time t; hence,

$$a_{i,t}(j) = P(i \xrightarrow{t} j) = P_{ij}(t).$$

Let $A_t \in M_{20}(R)$ be defined by

$$A_t e_i = (a_{i,t})^T,$$

so that A_t is the matrix with vectors $(a_{i,t})^T$ as columns, for all t. If we assume that A_t, in addition to (1) and (2), satisfies

$$P_{ij}(t + s) = \sum_k P_{ik}(t) P_{kj}(s), \forall s, t \in [0, \infty\rangle,$$

then A_t is the matrix of transition probabilities of a homogenous Markov process and can be written as

$$A_t = e^{tH},$$

describing the evolution of elements of A within the class P. There are several examples of such models in the context of biological sequence analysis, most notably the PAM series of matrices [39] - in the case of amino acid evolution - and Jukes-Cantor or Kimura matrices [40] in the case of DNA evolution. Now, we will present a simple substitution model, based on the BLOSUM matrices [16] - or, for that matter, on any substitution matrix - which does not necessarily arise from an evolutionary Markov process, but suffices for our purposes.

It is well known that the BLOSUM50 matrix is defined by $B(i, j) = \log_2 \frac{p_{ij}}{q_i q_j}$, $i, j \in \{1, \ldots, 20\}$, where $p_{i,j}$ indicates the probability of seeing amino acids i and j substitute each other in a homologous sequence. This matrix can also be written as

$$\frac{p_{ij}}{q_i} = e^{sB(i,j)} q_j,$$

for $s = \log_e 2$. Varying s in the above equation will, after renormalisation and reparametrisation $t = s^{-1}$ yield a family $a_{i,t}$ as above. This way of obtaining transition probabilities is clearly different and simpler than (5). However, it will produce a rich class of probability distributions that reflect relations between amino acids captured by BLOSUM scores.

Calculation of the evolutionary split statistic

In this section, we describe the *evolutionary split* (es) statistic. It will be used to predict positions in the multiple alignment that are potentially significant for functional clustering.

Definition 2 Let D denote a column in a multiple alignment, and assume that D contains no gaps. Then

$$es_k(D) = \log\left(\frac{\max_{\lambda_i, b_i} (P(D|\Sigma_1^k \lambda_i b_i))}{\max_b P(D|b)} \right),$$

where b, b_i are substitution distributions from Definition 1, $\sum_1^k \lambda_i = 1$, with $\lambda_i \geq 0$ and k is the number of subtypes that we are searching for. The algorithm was implemented as a C program.

Remark Note that $es_k(\cdot)$ compares the likelihood of the data with respect to the optimal mixture of k substitution models, with the likelihood under a single optimal model. In practice, we used a discrete approximation of the parameter space for the optimization. Also, a mild sequence weighting scheme was applied, to correct for the lack of independence in the sample (see [41]).

Clustering algorithm

Let us suppose that l columns (with no gaps) have been selected from the multiple alignment. Hence, we are dealing with n protein sequences $\gamma = \{\gamma^1, ..., \gamma^n\}$, all of the same length l, i.e $\gamma^i = (\gamma_1^i, \ldots, \gamma_l^i)$, for all i. We want to define a

model for dividing γ into k subsets. Let $I = (I_1, ..., I_k)$ stand for a partition of $\{1, ..., n\}$ into k non-empty disjoint subset. A model for our data set $\gamma = \{\gamma^1, ..., \gamma^n\}$ consists of two components -- a partition $I = (I_1, ..., I_k)$ and the parametric model M itself, which consists of k sequences of distributions from the substitution model, e.g. $(a^j_{i_1,t_1}, \ldots, a^j_{i_l,t_l})$, for $j = 1, ..., k$. We obtain the clustering by optimizing the following expression

$$\max_I \max_M \log P(\gamma \mid M, I),$$

where

$$P(\gamma \mid M, I) = \prod_{m=1}^{k} \prod_{i \in I_m} \prod_{j=1}^{l} P(\gamma^i_j \mid a^m_{i_j,t_j})$$

Thus, we rely on the conditional likelihood to cluster our data in k groups. By doing so, we effectively treat the partition $I = (I_1, ..., I_k)$ as a (discrete) parameter in the model.

A more traditional approach is to consider the real likelihood of the data with respect to the mixture model, and treat the membership of the clusters as missing data. In such a framework, the model M consists of parameters $\lambda_i \in [0,1)$, with $\sum_{i=1}^{k} \lambda_i = 1$ and k sequences of distributions from the substitution model as above. The model for the data is obtained by maximization of the log-likelihood

$$\max_M \log P(\gamma \mid M),$$

where

$$P(\gamma \mid M) = \prod_{i=1}^{n} \left(\sum_{m=1}^{k} \lambda_m \prod_{j=1}^{l} P(\gamma^i_j \mid a^m_{i_j,t_j}) \right)$$

Given the optimal model $M = \{(\lambda_i, M_i)\}$, we can obtain the clustering using the following Bayesian criterion

$$i \in I_j \Leftarrow \lambda_j P(\gamma^i \mid M_j) \geq \lambda_m P(\gamma^i \mid M_m), \textit{ for all } m.$$

Clearly the expression we need to optimize if we choose the conditional likelihood is much simpler, although the parameter space is somewhat more complicated. In either case, finding the optimal model is a difficult problem. For real-life data sets, the clustering will not differ if we choose one approach or the other, but the conditional likelihood procedure tends to reach the optimum much faster than the standard deterministic annealing EM-algorithm [42]. In some applications it might be more reasonable to take fully Bayesian approach and report posterior probabilities for each clustering obtained. However, our aim in the present paper was to obtain one useful partition of data sets and we did not explore this point of view further. In the rest of this section we describe a natural optimization method for the conditional likelihood approach.

Let us now describe the optimization algorithm. A clustering of the data set $\gamma = \{\gamma^1, ..., \gamma^n\}$ will be denoted by $I = (I_1, ..., I_k)$ -- same as the associated partition, and let M_i, $i = 1, ..., n$ denote the (parametric) model corresponding to the i-th cluster. As already mentioned, the following algorithm is a natural solution:

- Step1: choose an initial clustering $(I_1, ..., I_k)$
- Step2: determine the optimal model M_i for the i-th cluster, for all i
- Step3: for each γ^j, change cluster membership by setting $\gamma^j \in I_l$ if and only if $P(\gamma^j \mid M_l) \geq P(\gamma^j \mid M_i)$, for all i
- Step4: goto Step2

It is easy to show that this procedure increases the value of the likelihood function from (9), so will always reach a (local) maximum (if a sufficient number of iterations has been performed). In order to avoid local maxima, we use *smoothing*, i.e. we use the uniform distribution $u = (\frac{1}{20}, \ldots, \frac{1}{20})$ to obtain modified model $\hat{M}_i$ as a convex combination of M_j and u in Step2. Clearly, the amount of smoothing should be reduced as the optimization process progresses. Furthermore, we use simulated--annealing like acceptance-rejection principle for the cluster membership: the proposal in the Step3 is accepted with probability

$$\frac{kP(\gamma^j \mid M_l)}{T\Sigma_i P(\gamma^j \mid M_i)},$$

where T is the temperature, $+\infty \to T \to 0$. So, with these additions, we get the following algorithm:

- Step1: choose an initial clustering $(I_1, ..., I_k)$
- Step2: determine the optimal model M_i for the i-th cluster, for all i
- Step2': M_i is replaced with $\hat{M}_i$, for all i
- Step3': for each γ^j, propose cluster membership change by setting $\gamma^j \in I_l$ if and only if $P(\gamma^j \mid M_l) \geq P(\gamma^j \mid$

M_i), for all i, and accepting it with probability $\frac{k \cdot P(\gamma^j|M_l)}{T \cdot \sum_i P(\gamma^j|M_i)}$; if proposal is rejected, the cluster membership is assigned randomly

•Step4: goto Step2

The algorithm was implemented as a C program.

Availability

The programs are offered on a web server at: http://compbio.math.hr/. Further details of the programs can be obtained from PG.

Authors' contributions

PG developed the mathematical background of the clustering concept, produced the programs and wrote the initial draft of the manuscript. BB developed the statistical ideas. JZ, DV, AK and CE carried out the analyses of the protein families. DH, PFL and JC contributed biological ideas to the development of the methodology and drafted the final manuscript. All authors read and approved the final manuscript.

Additional material

Additional file 1

Alignments*. The alignments used to test the clustering method: nucleotidyl cyclases, protein kinases, dehydrogenases, acyl transferases, ketoreductases and small heat shock proteins.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-10-335-S1.TXT]

Additional file 2

Detailed output of the evolutionary split and clustering programs*. For each of the protein families (nucleotidyl cyclases, protein kinases, dehydrogenases, acyl transferases, ketoreductases and small heat shock proteins) there are two tables. The evolutionary split table has the following columns: residue position in the alignment (only residues that are present in every protein sequence are used for the calculation), amino acid for one ancestor model, log likelihood for one ancestor model, amino acids for two ancestor model, log likelihood for two ancestor model, evolutionary split statistic. The clustering table has the following columns: name of protein sequence, log likelihood for membership of subtype* ***a****, log likelihood for membership of subtype* ***b****, predicted subtype, specificity score.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-10-335-S2.XLS]

Acknowledgements

This work was mainly supported by the iProject 8045 M047 (to P.G.) and partially by the research grants 037-0982913-2762, 098-0982913-2877 and 058-0000000-3475 (to P.G., B.B., D.V. and D.H.) from the Ministry of Science, Education and Sports, Republic of Croatia. Additional support was received from the cooperation grant of the German Academic Exchange Service (DAAD) and the Ministry of Science, Education and Sports, Republic of Croatia (to J.C. and D.H.) and from the Leverhulme Trust, Japanese Bio-Industry Association and The School of Pharmacy, University of London (to PFL). A.K. is grateful to UNESCO and L'Oreal for the fellowship in the framework of the program "For Women in Science".

References

1. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215:**403-410.
2. Eddy SR: **Profile hidden Markov models.** *Bioinformatics* 1998, **14:**755-763.
3. Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer EL: **The Pfam protein families database.** *Nucleic Acids Res* 2002, **30:**276-280.
4. Hranueli D, Cullum J, Basrak B, Goldstein P, Long PF: **Plasticity of the *Streptomyces* genome - evolution and engineering of new antibiotics.** *Curr Med Chem* 2005, **12:**1697-1704.
5. Chan YA, Podevels AM, Kevany BM, Thomas MG: **Biosynthesis of polyketide synthase extender units.** *Nat Prod Rep* 2009, **26:**90-114.
6. Starcevic A, Zucko J, Simunkovic J, Long PF, Cullum J, Hranueli D: ***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* 2008, **36:**6882-6892.
7. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** *Nucleic Acids Res* 1994, **22:**4673-4680.
8. Caffrey P: **Conserved amino acid residues correlating with ketoreductase stereospecificity in modular polyketide synthases.** *Chem Bio Chem* 2003, **4:**654-657.
9. Yadav G, Gokhale RS, Mohanty D: **Computational approach for prediction of domain organization and substrate specificity of modular polyketide synthases.** *J Mol Biol* 2003, **328:**335-363.
10. Hannenhalli SS, Russell RB: **Analysis and prediction of functional sub-types from protein sequence alignments.** *J Mol Biol* 2000, **303:**61-76.
11. Pirovano W, Feenstra KA, Heringa J: **Sequence comparison by sequence harmony identifies subtype-specific functional sites.** *Nucleic Acids Res* 2006, **34:**6540-6548.
12. Pazos F, Rausell A, Valencia A: **Phylogeny-independent detection of functional residues.** *Bioinformatics* 2006, **22:**1440-1448.
13. Wallace IM, Higgins DG: **Supervised multivariate analysis of sequence groups to identify specificity determining residues.** *BMC Bioinformatics* 2007, **8:**135.
14. Ye KK, Feenstra A, Heringa J, IJzerman AP, Marchiori E: **Multi-RELIEF: a method to recognize specificity determining residues from multiple sequence alignments using a machine-learning approach for feature weighting.** *Bioinformatics* 2008, **24:**18-25.
15. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG: **The CLUSTAL X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools.** *Nucleic Acids Res* 1997, **25:**4876-4882.
16. Henikoff S, Henikoff JG: **Amino acid substitution matrices from protein blocks.** *Proc Natl Acad Sci USA* 1992, **89:**10915-10919.
17. The UniProt Consortium: **The Universal Protein Resource (UniProt) 2009.** *Nucleic Acids Res* 2009, **37:**D169-D174.
18. Smith CM, Shindyalov IN, Veretnik S, Gribskov M, Taylor SS, Ten Eyck LF, Bourne PE: **The protein kinase resource.** *Trends Biochem Sci* 1997, **22:**444-446.
19. Wilks HM, Hart KW, Feeney R, Dunn CR, Muirhead H, Chia WN, Barstow DA, Atkinson T, Clarke AR, Holbrook JJ: **A specific, highly acitve malate dehydrogenase by redesign of a lactate dehydrogenase framework.** *Science* 1988, **242:**1541-1544.
20. Haydock SF, Aparicio JF, Molnár I, Schwecke T, Khaw LE, König A, Marsden AF, Galloway IS, Staunton J, Leadlay PF: **Divergent sequence motifs correlated with the substrate specificity of (methyl)malonyl-CoA:acyl carrier protein transacylase domains in modular polyketide synthases.** *FEBS Lett* 1995, **374:**246-248.

21. Lau J, Fu H, Cane DE, Khosla C: **Dissecting the role of acyltransferase domains of modular polyketide synthases in the choice and stereochemical fate of extender units.** *Biochemistry* 1999, **38:**1643-1651.
22. Reeves CD, Murli S, Ashley GW, Piagentini M, Hutchinson CR, McDaniel R: **Alteration of the substrate specificity of a modular polyketide synthase acyltransferase domain through site-specific mutations.** *Biochemistry* 2001, **40(51):**15464-15470.
23. Del Vecchio F, Petkovic H, Kendrew SG, Low L, Wilkinson B, Lill R, Cortés J, Rudd BA, Staunton J, Leadlay PF: **Active-site residue, domain and module swaps in modular polyketide synthases.** *J Ind Microbiol Biotechnol* 2003, **30:**489-494.
24. Serre L, Verbree EC, Dauter Z, Stuitje AR, Derewenda ZS: **The *Escherichia coli* malonyl-CoA:acyl carrier protein transacylase at 1.5A resolution. Crystal structure of a FAS component.** *J Biol Chem* 1995, **270:**12961-12964.
25. Castonguay R, He W, Chen AY, Khosla C, Cane DE: **Stereospecificity of ketoreductase domains of the 6-deoxyerythronolide B synthase.** *J Am Chem Soc* 2007, **129:**13758-13769.
26. Waters ER, Lee GJ, Vierling E: **Evolution, structure and function of the small heat shock proteins in plants.** *J Exp Bot* 1996, **47:**325-338.
27. van Montfort RL, Basha E, Friedrich KL, Slingsby C, Vierling E: **Crystal structure and assembly of a eukaryotic small heat shock protein.** *Nat Struct Biol* 2001, **8:**1025-1030.
28. Kim KK, Kim R, Kim SH: **Crystal structure of a small heat-shock protein.** *Nature* 1998, **394:**595-599.
29. Starcevic A, Jaspars M, Cullum J, Hranueli D, Long PF: **Predicting the nature and timing of epimerisation on a modular polyketide synthase.** *Chem Bio Chem* 2007, **8:**28-31.
30. Keatinge-Clay AT: **A tylosin ketoreductase reveals how chirality is determined in polyketides.** *Chemistry & Biology* 2007, **14:**898-908.
31. Veerassamy S, Smith A, Tillier ERM: **A transition probability model for amino acid substitutions from blocks.** *J Comput Biol* 2003, **10:**997-1010.
32. Jones DT, Taylor WR, Thornton JM: **The rapid generation of mutation data matrices from protein sequences.** *Comput Appl Biosci* 1992, **8:**275-282.
33. **ExPASy Proteomics Server** [http://expasy.org/]
34. **NRPS_PKS: A knowledge based resource for analysis of Non-ribosomal Peptide Synthetases and Polyketide Synthases** [http://www.nii.res.in/nrps-pks.html]
35. Ansari MZ, Yadav G, Gokhale RS, Mohanty D: **NRPS-PKS: a knowledge-based resource for analysis of NRPS/PKS megasynthases.** *Nucleic Acids Res* 2004, **32(Web server issue):**W405-W413.
36. Tae H, Jae KS, Park K: **Development of an analysis program of Type I polyketide synthase gene clusters using homology search and profile hidden Markov model.** *J Microbiol Biotechnol* 2009, **19:**140-146.
37. **European Bioinformatics Institute** [http://www.ebi.ac.uk]
38. Felsenstein J: **PHYLIP - Phylogeny Inference Package (Version 3.2).** *Cladistics* 1989, **5:**164-166.
39. Dayhoff MO, Schwartz RM, Orcutt BC: **A model of evolutionary change in proteins.** *Atlas of Protein Sequence and Structure* 1978, **5:**345-352.
40. Felsenstein J: **Inferring Phylogenies.** Sunderland, MA: Sinauer Associates; 2004.
41. Henikoff S, Henikoff JG: **Position-based sequence weights.** *J Mol Biol* 1994, **243:**574-578.
42. Ueda N, Nakano R: **Deterministic Annealing EM Algorithm.** *Neural Networks* 1998, **2:**271-282.

2.2 *ClustScan: An integrated program package for the semi-automatic annotation of modular biosynthetic gene clusters and in silico prediction of novel chemical structures*

Antonio Starcevic, **Jurica Zucko**, Jurica Simunkovic, Paul F. Long, John Cullum and Daslav Hranueli. *Nucleic Acids Res.*, ***36***, 6882-6892, 2008.

Abstract:

The program package ***ClustScan*** is designed for rapid, semi-automatic, annotation of DNA sequences encoding modular biosynthetic enzymes including polyketide synthases (PKS), non-ribosomal peptide synthetases (NRPS) and hybrid (PKS/NRPS) enzymes. The program displays the predicted chemical structures of products as well as allowing export of the structures in a standard format for analyses with other programs. Recent advances in understanding of enzyme function are incorporated to make knowledge-based predictions about the stereochemistry of products. The program structure allows easy incorporation of additional knowledge about domain specificities and function. The results of analyses are presented to the user in a graphical interface, which also allows easy editing of the predictions to incorporate user experience. The versatility of this program package has been demonstrated by annotating biochemical pathways in microbial, invertebrate animal and metagenomic datasets. The speed and convenience of the package allows the annotation of all PKS and NRPS clusters in a complete *Actinobacteria* genome in 2-3 man hours. The open architecture of ***ClustScan*** allows easy integration with other programs, facilitating further analyses of results, which is useful for a broad range of researchers in the chemical and biological sciences.

Own contribution to the paper:

Developing methods for determining substrate specificity of AT domains and activity of ER and DH domains.

6882–6892 *Nucleic Acids Research, 2008, Vol. 36, No. 21* *Published online 31 October 2008*
doi:10.1093/nar/gkn685

ClustScan: an integrated program package for the semi-automatic annotation of modular biosynthetic gene clusters and *in silico* prediction of novel chemical structures

Antonio Starcevic[1,2], Jurica Zucko[2,3], Jurica Simunkovic[3], Paul F. Long[4], John Cullum[2] and Daslav Hranueli[1,*]

[1]Faculty of Food Technology and Biotechnology, University of Zagreb, Pierottijeva 6, 10000 Zagreb, Croatia, [2]Department of Genetics, University of Kaiserslautern, Postfach 3049, 67653 Kaiserslautern, Germany, [3]Novalis Ltd, Božidara Adžije 17, 10000 Zagreb, Croatia and [4]The School of Pharmacy, University of London, 29/39 Brunswick Square, London WC1N 1AX, UK

Received June 19, 2008; Revised September 23, 2008; Accepted September 24, 2008

ABSTRACT

The program package '*ClustScan*' (*Clust*er *Scan*ner) is designed for rapid, semi-automatic, annotation of DNA sequences encoding modular biosynthetic enzymes including polyketide synthases (PKS), non-ribosomal peptide synthetases (NRPS) and hybrid (PKS/NRPS) enzymes. The program displays the predicted chemical structures of products as well as allowing export of the structures in a standard format for analyses with other programs. Recent advances in understanding of enzyme function are incorporated to make knowledge-based predictions about the stereochemistry of products. The program structure allows easy incorporation of additional knowledge about domain specificities and function. The results of analyses are presented to the user in a graphical interface, which also allows easy editing of the predictions to incorporate user experience. The versatility of this program package has been demonstrated by annotating biochemical pathways in microbial, invertebrate animal and metagenomic datasets. The speed and convenience of the package allows the annotation of all PKS and NRPS clusters in a complete *Actinobacteria* genome in 2–3 man hours. The open architecture of *ClustScan* allows easy integration with other programs, facilitating further analyses of results, which is useful for a broad range of researchers in the chemical and biological sciences.

INTRODUCTION

Bioprospecting for lead compounds from nature continues to be a corner stone in drug development. As well as isolating microorganisms from unique environments or biological diversity 'hotspots', approaches are also being developed to exploit the chemical diversity from > 98% of uncultivable microbes living in the natural environment. There is now unprecedented opportunity to access the natural diversity of small molecules made by such microbes by the isolation of metagenomic DNA and heterologous expression of biosynthetic pathways in a fermentable host. Discovery of novel biosynthetic gene clusters is the first goal of this culture-independent research that requires the application of molecular bioinformatics to identify DNA sequences of interest. We have developed an integrated set of computer programs for this task, which we call the '*ClustScan*' (*Clust*er *Scan*ner) program package.

Many important secondary metabolites in bacteria are synthesized on enzymes encoded by modular biosynthetic gene clusters: polyketide synthase (PKS) clusters, non-ribosomal peptide synthetase (NRPS) clusters, NRPS-independent siderophore (NIS) synthetase clusters or hybrid clusters (1–4). These secondary metabolites include polyketide antibiotics (e.g. erythromycin), immuno-suppressants (e.g. rapamycin) and antiparasitics (e.g. avermectin) as well as peptide antibiotics (e.g. vancomycin), immuno-suppressants (e.g. cyclosporin) and herbicides (e.g. bialaphos). Correlation of the chemical structures of the products with cluster DNA sequences shows that, in most cases, a defined series of catalytic domains that can be grouped into modules are responsible

*To whom correspondence should be addressed. Tel: +385 1 4605013, Fax: +385 1 4836083; Email: dhranueli@pbf.hr

for each round of chain elongation. Thus, synthesis follows a co-linear principle in which the gene sequences of the individual modules determine the chemical outcomes of successive chain extension reactions. Large-scale DNA sequencing has revealed many gene clusters, whose products are not known (5–7). Predictions about the structures of the products based on the DNA sequences encoding enzyme modules can help decisions about which products may be interesting in the search for novel drugs. Modules are composed of domains that carry out the different reactions so that prediction of module specificity can be built up from that of domain specificity. In PKSs, each module usually contains an acyl carrier protein (ACP) domain and an acyl transferase (AT) domain, which is responsible for substrate selection and transferring the substrate to the ACP domain. For all modules except possibly the starter module ('loading domain') there is also a ketosynthase (KS) domain that performs condensation. Some AT domains select a malonyl–CoA substrate which results in a two carbon extension. However, other substrates can be used (e.g. methylmalonyl–CoA, ethylmalonyl–CoA, methoxymalonyl–CoA) which result in the incorporation of more carbon atoms. However, the backbone chain is always extended by two carbon atoms and the other carbon atoms occur as side chains (e.g. methyl groups). Amino acid residues in AT domains that differ between malonyl–CoA-incorporating and methylmalonyl–CoA-incorporating have been identified from multiple alignments of AT sequences (8–12). There may be further reduction domains that carry out a sequential reduction of the introduced keto group: ketoreductase (KR) produces a hydroxyl group, which may be acted on by a dehydratase domain (DH) to produce a double bond that can be modified to a completely reduced product by an enoyl reductase (ER) domain. The stereochemistry of the addition step is also important. This can arise when the KR domain introduces a hydroxyl group and comparison of the sequences of KR domains introducing different stereochemistry identified specific residues correlated with this difference (13,14). A second source of differential stereochemistry is the incorporation of an extender unit with more than two carbon atoms resulting in a side chain with a choice of stereochemistry. At one time it was assumed that the KS domain was responsible for this choice. However, bioinformatic analyses could find no amino acid differences in the KS domain correlating with the stereochemical outcome and instead found correlations with the sequence of the KR domain (15). Studies of the 3D structure of KR domains provided mechanistic explanations of how the stereochemistry of the hydroxyl group and the α-carbon atom are controlled (16,17). The chirality of the α-carbon is lost if reduction of the hydroxyl group to a double bond on the β-carbon occurs, but this reduction may result in a new stereochemical choice between the *cis*- and *trans*-isomers that is probably determined by the DH domain carrying out the reduction. A new chirality may be created if full reduction occurs and is likely to be determined by the ER domain responsible for the final reduction step.

The annotation of the DNA sequence of a PKS cluster can be time-consuming because of the large number of domains and the necessity of integrating data from many sources. Several tools have been developed to assist this process. Identifying domains poses few problems as the sequences are well conserved. A much more difficult problem is predicting the activity and specificity of domains. The NRPS–PKS database (18, http://www.nii.res.in/nrps-pks.html), holds data on PKS and NRPS gene clusters including module and domain structure and chemical structures of the biosynthetic products. It allows users to input protein sequences to be used in BLAST (19) searches to identify domains and finds the closest sequences in the database. This allows prediction 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). The ASMPKS database (20, http://gate.smallsoft.co.kr:8008/%7Ehstae/asmpks/index.html) uses a similar methodology, but integrates it with a graphical display of the domains in genes so that modules can be easily recognized. It also allows the display of a predicted linear polyketide chain product for which the user has to select starter and extender units from lists. Minowa *et al.* (21) used an approach based on the creation of hidden Markov model profiles (22) to predict substrate specificity of AT domains. The company ECOPIA has also developed a software tool (23) DecipherIT^{TM}, which helps annotation of new gene clusters based on comparison with a database of known clusters. Although these approaches are useful, they do not make predictions about the stereochemistry of the products, which is extremely important for assessing their promise. As these analyses are essentially based on similarity to known clusters rather than identification of functional residues, they are less effective for clusters from novel organism groups. Another practical limitation is that they do not export information about chemical structures in a format that can be used by standard programs for further analyses.

In this paper, we describe a program that utilizes recent advances in understanding the function of KR domains to make knowledge-based predictions of activity and stereochemical specificity for hydroxyl groups and α-carbon atoms. This is combined with a fingerprint approach to predict specificity of AT domains and more conventional approaches for prediction of activity of DH and ER domains. The program predicts the chemical structures of products, which can be exported in a SMILES/SMARTS format for further analysis by standard Chemistry programs. The program is structured so that it can easily be updated to incorporate new knowledge about the specificity of domains. It has a convenient graphical interface that allows the rapid semi-automatic annotation of gene clusters encoding modular biosynthetic enzymes by non-expert users.

MATERIALS AND METHODS

GeneMark (24) (version 2.5; http://opal.biology.gatech.edu/GeneMark/) or Glimmer (25) (version 3.02; http://www.cbcb.umd.edu/software/glimmer/) were used to identify genes. HMMER (22) (version 2.3.2; http://hmmer.janelia.org/) was used for identification of

6884 Nucleic Acids Research, 2008, Vol. 36, No. 21

protein domains. Profiles from Pfam (26) as well as specially constructed profiles were used. The gene prediction and protein domain prediction programs run on a Linux server and each user has a password to allow access to their own workspace. All user activities are performed via the Java client, which was written in Windows, MacIntosh and Linux versions.

To predict the specificity of AT or KR domains the amino acid sequence was aligned with an appropriate HMMER profile and the diagnostic amino acid residues extracted (Supplementary Data 1 Tables 2S and 3S). The diagnostic residues were compared to fingerprints corresponding to the different specificities (substrate specificity for AT; activity and stereochemistry for KR). The prediction of activity/inactivity of DH domains used a HMMER-profile based on active actinomycete domains. The prediction was based on the HMMER score. ER domains were detected using a profile based on a mixture of active and inactive domains.

To predict chemical structures, a table was constructed (see Supplementary Data 1 Table 4S) that contained different chemical building blocks written as isomeric SMILES (27). These were ordered on the basis of substrate and degree of reduction. In cases, where stereochemical prediction was not possible non-isomeric SMILES were used. Generic units as SMARTS (http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html) were also included for cases where prediction was not possible. The predictions were used to generate a description of the product in an XML format (http://www.w3.org/TR/xml/) organized in a hierarchical structure corresponding to module and domain architecture. This XML description was used to generate the chemical structure from the table of SMILES. This description was also used to generate a ring structure from the linear polyketide using a simple cyclization rule. The SMILES description can be drawn and displayed in *ClustScan* using Jmol v. 11.2.14, 2006 (Jmol; http://www.jmol.org/) or exported. *Clustscan* can be obtained by request from Novalis Ltd (novalis@novalis.hr).

RESULTS

The analyses of the DNA sequence data are carried out on a server and the results are cached so that each analysis only needs to be carried out once. This is important as the analysis of a whole *Streptomyces* genome may take several hours, but this can occur unsupervised overnight. The user accesses the results using a Java client that gives user-friendly presentation of the data. There is a password-protected workspace for each user on the server. The client allows the user to upload DNA sequences to the server and initiate analyses. The sequence is automatically translated in all six reading frames to allow HMMER (22) searches using a library of protein family profiles. The standard libraries contain PKS and/or NRPS domains, but it is possible to add other profiles if desired that makes the program package generic. These can be profiles from the Pfam (26) database or custom profiles created with the HMMER package (e.g. we have used profiles to find and annotate shikimic acid pathway genes; see Performance of *ClustScan* subsection). Independently of the search for protein patterns, the DNA can be analyzed to find probable coding regions using GeneMark (24) or Glimmer (25). GeneMark provides a library of models based on different bacteria and the appropriate model is chosen using a species related to the source of the DNA. Glimmer can construct a model for coding regions using long open reading frames (ORF) in the input sequence as training data. This is less effective for short input sequences. Also sequences with high G + C-content have long non-coding random ORFs, which may reduce the accuracy of coding sequence prediction. The program, therefore, also allows the user to create a model by supplying appropriate training data (e.g. the genome sequence of a related species) and the model can be stored by the user for future analyses.

The results of the analyses are presented both as lists in the 'workspace' window (Figure 1A) and graphically in the 'annotation' editor window (Figure 1B). The workspace window shows the results in a tree format in which branches can be opened up or collapsed to show the genes and the protein domains. This is useful for obtaining an overview and it is possible to navigate through the thousands of genes present in a complete genome. The graphical 'annotation' editor window (Figure 1B) shows the positions of genes and protein domains on the six reading frames and can be viewed at different resolutions using a zoom function. It is possible using the mouse to displace genes and domains above and below the reading frames for better visualization of overlapping regions. It is usual to keep both the workspace window and the annotation editor window open and clicking on a feature in either, marks the corresponding feature in the other window. The protein domains are identified by HMMER analysis using a cut-off score that can be set to a stringent or relaxed value. This results in some putative protein domains, which may not be genuine. The user can choose to reject a protein domain so that it is removed from the analysis; the program tracks editing changes so that they can later be reversed if mistaken deletions occur. In many cases, the decision about the protein domain is taken on the basis of whether it occurs at an appropriate position with respect to other domains, which is easily seen in the graphical view of the annotation editor. To help the decision, the evidence for the identification of the protein domains can be viewed using the 'details' window (Figure 2). This shows the coordinates of the protein domain in the DNA and protein and the scores and E-values from the HMMER analysis. In addition, the alignment of the protein with the profile is shown. A prediction of the specificity of the protein domain is also shown. For AT domains (Figure 2A) this is the starter unit incorporated by the condensation reaction. For KR domains (Figure 2B) it is predicted whether the domain is active for reduction and, in addition, the stereochemistry of the hydroxyl group and the α-carbon atom are predicted. The predictions can be overridden if the user has extra information in conflict with the program's prediction. For instance, Figure 2A shows the (correct) prediction of propionyl as the starter unit

Figure 1. (**A**) The workspace window gives an overview of the analysis in the form of collapsible trees. Detected genes and protein domains are shown. (**B**) The annotation editor window shows the location of genes (in red) and protein domains (in blue). In this case there are three genes on the three different forward open reading frames. The genes have been displaced from the reading frames by the user to allow better visualization of the domains. The annotation editor has been used for user definition of modules (shown as red curves below the open reading frames). (**C**) The cluster editor window. The user can define a set of contiguous genes as a cluster. The cluster editor window shows the genes in a cartoon form with an expanded view of the selected gene showing protein domains. Domains can be linked together to give modules. The modules are given identifying names and the program suggests a biosynthetic order that can be accepted or altered by the user.

for erythromycin. By clicking on the propionyl, a drop-down list is shown that enables selection of an alternative unit.

On the basis of the results in the annotation editor, the user can define a gene cluster covering a region of adjacent genes. The annotation of the gene cluster is carried out using the 'cluster' editor window (Figure 1C), which shows the genes of the gene cluster in a simple cartoon form, hence the term semi-automatic. When a gene is selected, the protein domains are also shown. Modules can be assembled by marking protein domains and each module created is given a name. The program suggests a biosynthetic order of the genes of a cluster. For PKS clusters this is based on identifying a potential loading domain (i.e. typically a module containing only AT and ACP domains; Figure 1C) and looking for a thioesterase domain as identifying the last module. If there is ambiguity, it is assumed that the genes are used in the pathway in the same order as they occur in the DNA. This procedure identifies the correct biosynthetic order in most natural gene clusters. The user can alter the suggested order to incorporate any additional knowledge available.

The complete annotation by *ClustScan* can be stored as a file in an XML format so that it can be reimported into *ClustScan*. The hierarchical nature of XML makes it well suited for representing clusters in terms of genes, modules and protein domains. We developed an XML format that includes information about the biosynthetic order. Although the XML format is primarily designed for the internal use of *ClustScan*, it makes it easy for other applications to read or write *ClustScan* compatible files by adding an appropriate XML parser. In addition to the XML format, annotations can be exported as an EMBL or GenBank file for use in other applications or for submission to databases; this results in loss of information on biosynthetic order. In addition, the DNA or amino acid sequences of genes, domains or modules can be copied to the clipboard for further analyses with other programs.

The prediction functions for the activity and specificity of protein domains are used to deduce module specificity and, thus, to predict the chemical structure of the linear polyketide chain product of the gene cluster. The structures are represented internally in the program as isomeric SMILES (27), which can be copied to the clipboard (Figure 3A) allowing export for use with standard

6886 Nucleic Acids Research, 2008, Vol. 36, No. 21

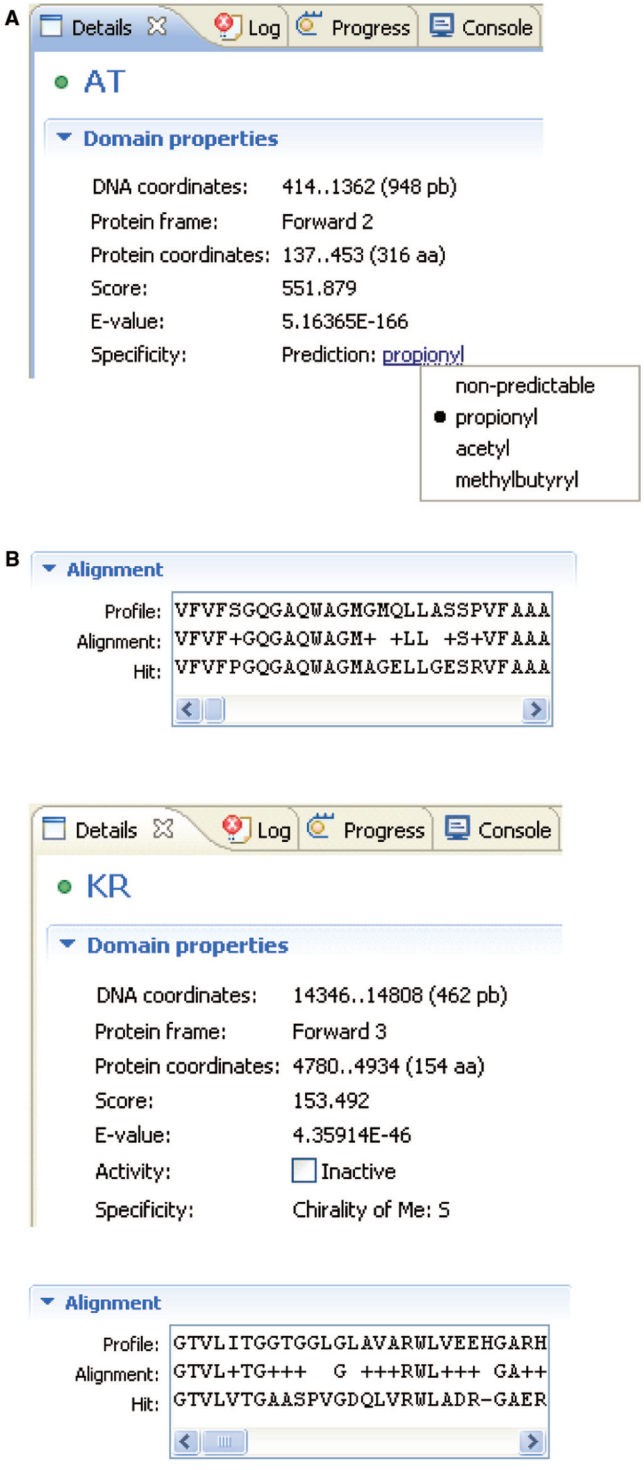

Figure 2. The details window allows the user to examine the evidence for assignment of protein domains. The HMMER scores and E-values as well as the alignment are displayed. The predictions of activity and specificity are also displayed and can be modified by the user. (**A**) The loading AT domain of the erythromycin cluster. The program makes the correct prediction of a propionyl starter unit. By clicking on this choice, a selection window has been opened that allows the user to override the automatic prediction and select an alternative choice. (**B**) The KR domain of module 3 of the erythromycin cluster.

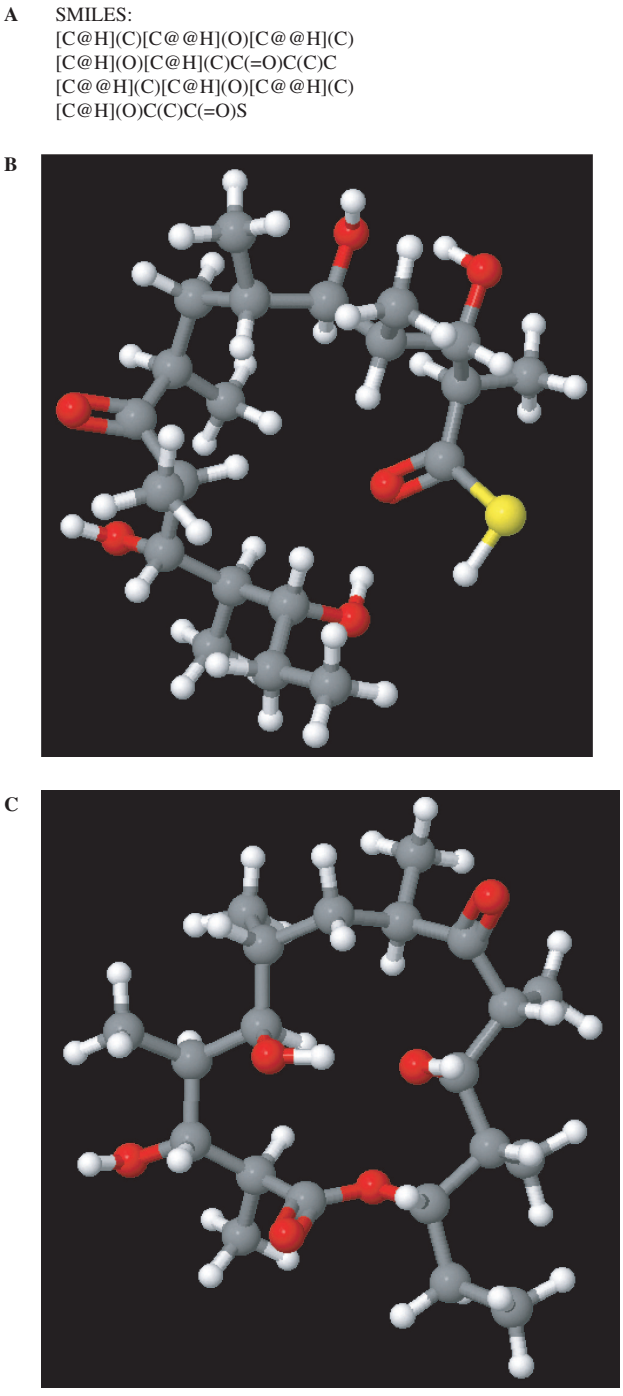

Figure 3. The molecules window. (**A**) The SMILES description for the linear backbone of erythromycin predicted from the DNA sequence of the cluster. The SMILES description can be copied to the clipboard for export. (**B**) The 3D structure of the predicted linear chain is shown. The mouse can be used to rotate the molecule. (**C**) The ring structure of the erythromycin aglycone as predicted using the cyclization function of the program.

chemical software. The user can define new module specificities and provide isomeric SMILES descriptions of the extender units. It is possible to edit the prediction of module specificity to allow incorporation of such novel

extender units. The program allows the user to display the chemical structure of products in a 3D 'molecules' window (Figure 3B) in which the molecule can be rotated. The program can also produce a potential cyclic structure from a linear molecule (Figure 3C). It is assumed that the first hydroxyl or amino group introduced during synthesis reacts with the terminal extender unit.

The program is designed to allow easy incorporation of new knowledge. New or modified prediction of enzyme activity or specificity can be implemented without changing program structure. It is also possible for sophisticated users to write their own specific scripts to introduce specialized prediction functions.

Prediction of domain activities

The presence of any of the seven domains KS, AT, ER, DH, KR, ACP or TE is detected using the HMMER profiles. An extender module needs at least KS, AT and ACP. AT determines the substrate selection for the extension reaction. The three reduction domains (KR, DH and ER) may be absent or present as active or inactive domains. *ClustScan* predicts whether domains are active as well as predicting substrate specificity or stereochemical outcome when several outcomes are possible (see Supplementary Data 1 Table 1S).

The KR domain is the best characterized domain in terms of structural determination of differential activities. Active KR domains determine the chirality of the hydroxyl product and bioinformatic analysis identified amino acid residues involved in this choice (13,14). Bioinformatics also suggested that KR rather than KS determined the stereochemistry of β-carbon groups, when C3 or C4 units are incorporated (15). A comparison of 3D structures of two KR domains of different specificity gave more detailed information on amino acid residues involved in determining both hydroxyl and β-carbon stereochemistry (17). In *ClustScan*, alignment with a KR profile allows identification of all of these critical amino acids (the 'fingerprint') and, thus prediction of the product. The fingerprints used are shown in the Supplementary Data 1 Table 2S. There are six possible products (A1, A2, B1, B2, C1, C2), which correspond to three possible ketoreduction outcomes (either hydroxyl stereoisomer—A or B, or no reduction C) coupled with two β-carbon chiralities (called 1 and 2). The accuracy of prediction was tested using 49 KR domains for which the structure of the polyketide product provides information about activity and stereochemistry; if further active reduction domains are present, the product does not provide any information about the stereochemistry of the KR step. Ten of the KR domains processed 2-carbon extender units so that only hydroxyl stereochemistry was relevant: all 10 predictions were correct. Nine of the KR domains processing 3- or 4-carbon extender units were inactive: in eight cases the program predicted that the domains were inactive for reduction and also predicted the correct side chain stereochemistry. In one case the inactive KR domain was predicted as active. The other 30 KR domains processing 3- or 4-carbon extender units were active. In 25 cases the program predicted the correct stereochemistry. In one case, the program predicted the incorrect side chain stereochemistry (A1 instead of A2). In the other four cases, the alignment with the profile did not yield an amino acid fingerprint that fell into any of the groups: in these cases the program indicates that no prediction is possible. Thus, the KR prediction was correct in 88% of the cases, incorrect in 4% of the cases and the program was unable to provide a prediction in 8% of the cases.

Unlike the case of KR, structural information about AT domains is not sufficient to help in substrate prediction. The most common extender substrates are malonyl–CoA and methylmalonyl–CoA. Comparison by eye of alignments of AT domain sequences identified 13 amino acid residues, which differed significantly between domains incorporating the two substrates (8–12). The amino acid sequences of nine AT extender domains that incorporated ethylmalonyl–CoA were examined. It was found that the 13 amino acid residues had a common pattern that differed from those of the malonyl–CoA and methylmalonyl–CoA-specific AT domains. This information was used for prediction of specificity in the program. A further known extender substrate is methoxymalonyl–CoA and specific residues associated with choice of this substrate were identified in AT domains of the concanamycin A cluster (28). Eleven methoxymalonyl-incorporating AT domains were examined, but the 13 fingerprint residues used to characterize the other substrates did not show a conserved pattern. It was noticed that most had insertions with respect to the conserved alignment of all AT domains, which caused problems in identifying potential fingerprinting residues. After using a specific alignment for methoxymalonyl–CoA-incorporating AT domains, it was possible to use a modified form of the published pattern (28) to predict methoxymalonyl–CoA as a substrate.

The information about AT extender specificity was implemented in *ClustScan*. The amino acid sequence of the AT domain was aligned with a general AT-profile to identify the 13 diagnostic amino acid residues. These were compared to three fingerprints corresponding to the three substrates. If the amino acids did not fit any of the three fingerprints, the AT domain was aligned using a profile derived from the 11 methoxymalonyl–CoA AT domains. This alignment was used to test if one of the characterized insertions was present. If no match was found, the AT domain was assigned to an unknown substrate category.

In addition to AT domains in extender modules, there are often AT domains in loading domains. A set of AT domains that incorporate acetyl starters (nine domains), propionyl starters (eight domains) or methylbutyryl starters (three domains) were aligned with the general AT profile and the 13 diagnostic amino acid residues extracted. The fingerprints for acetyl and propionyl starters were identical to those for acetyl and propionyl extenders, respectively. The methylbutyryl starters showed a different pattern, which was also used to construct a specific fingerprint. This information was incorporated into *ClustScan*. When the user accepts the suggested biosynthetic order or defines a different order the loading domain is subjected to a special analysis. If an AT domain is

present, the 13 diagnostic amino acids are extracted and tested for acetyl, propionyl or methylbutyryl fingerprints. All fingerprints for the specificity of AT starter and extender units are shown in Supplementary Data 1 Table 3S. A dataset of 196 known AT domains was analyzed with *ClustScan* (95 malonyl–CoA, 79 methylmalonyl–CoA, 9 ethylmalonyl–CoA and 13 methoxymalonyl–CoA. The remaining 25 were propionyl, acetyl, methylbutyryl and some unusual ones from loading domain ATs). This gave the correct prediction in 182 cases (93%), the wrong prediction in 9 cases (5%) and assignment to an unknown class in 5 cases (3%).

For DH domains, the prediction should distinguish active and inactive domains. As insufficient structural information was available to make predictions based on knowledge of function, it was decided to use a profile based on active domains to try to predict activity. The profile was built using the sequences of 57 active domains derived from actinomycetes. The profile was used to screen the active domains used in its construction as well as an additional 56 active and 46 inactive domains (159 total). All domains with a high score (>300) were active, whereas all with a low score (<200) were inactive. About 80% of the domains with intermediate scores were active, but the scores of inactive domains were distributed through the range. These results were used to define a prediction function with three outcomes: active (score >300), 80% probability of activity (scores between 200 and 300) and inactive (score < 200). This prediction function was tested on 159 domains (113 active and 46 inactive). Forty-six domains fell into the intermediate region (36 active, 10 inactive) with a prediction of 80% probability of activity. Sixty-seven domains were predicted to be active of which six were in fact inactive corresponding to a 9% false prediction rate. In contrast, the prediction of inactivity was less satisfactory: 43 DH domains were predicted to be inactive of which 16 were actually active. A closer examination of these false predictions showed that only 1 of the 16 was an actinomycete sequence, the other 15 being sequences from Gram-negative bacteria. When attention was confined to actinomycete sequences, 13 DH domains were predicted to be inactive of which only one was active.

Initially, a similar approach to that used for the DH domain was attempted with the ER domains. A profile was constructed using active actinomycete ER domains. However, it was found that better prediction was achieved with a profile based on a mixture of active and inactive domains. Sixty-six known ER domains were tested. In all cases the ER domain was detected. The HMMER score did not prove useful in distinguishing between active and inactive ER domains. However, there were only three cases of inactive ER domains in the presence of an active DH domain. There were four cases in which an ER domain was detected, but the DH domain was inactive. The program, therefore, predicts an active ER domain if a domain is found and there are active KR and DH domains present. This gives a false prediction in the 3/66 (5%) cases of an inactive ER domain with an active DH domain.

Performance of *ClustScan*

There are two main criteria for the usefulness of *ClustScan*: the accuracy of prediction and the speed and convenience of annotating large datasets. The accuracy of prediction was tested on two well-known gene clusters: the erythromycin gene cluster and the niddamycin gene cluster (GenBank accession numbers AY771999 and AF016585). For the erythromycin gene cluster, with one exception, all the protein domains of the six extender modules were accurately identified and the propionyl starter (Figure 2A) was also predicted. The only exception was that *ClustScan* was not able to predict the hydroxyl group stereochemistry of the KR domain of module 4; the prediction of the hydroxyl stereochemistry is flagged as unknown. This does not have an effect on the final prediction as an active DH domain forms a double bond. However, the active ER domain recreates a chiral center, which cannot be predicted with the current state of knowledge. This resulted in two possible structures, where the user can choose the correct chirality to obtain an accurate prediction of the chemistry of the linear backbone (Figure 3A and B). In this case, the cyclization was also predicted correctly (Figure 3C) (see also Supplementary Data 1 Figure 1S A and B). In the niddamycin gene cluster, the five genes, the loading domain and the seven extender modules containing 36 catalytically active domains were all correctly predicted with the exception that the substrate for module six was predicted as ethylmalonate instead of the correct methoxymalonate. The inactive KR in module 4 responsible for the β-carbon: S stereochemistry was predicted. The correct cyclization was also predicted (see Supplementary Data 1 Figure 2S A and B). The results with *ClustScan* were compared with those from the NRPS–PKS database prediction system (SEARCHPKS), which is the most popular current analysis tool for PKS clusters (see Supplementary Data 2 Figures 1–4). SEARCHPKS (http://www.nii.res.in/nrps-pks.html) requires protein sequences so the amino acid sequences of the genes were extracted with *ClustScan* and submitted. SEARCHPKS found two extra false positive ACP domains in the erythromycin cluster (Supplementary Data 2 Figure 1). The first at the end of the *eryAI* gene did not affect the prediction as it was an isolated ACP domain. The second occurred between the KS and AT domains of module five and resulted in the program predicting an additional module and making no prediction of the chemistry of the two modules generated. It is not possible to review the data behind the prediction or to manually reject the false positives. SEARCHPKS found all the other domains successfully, but does not attempt to make predictions of the activity or stereochemistry of the reduction domains. In particular, this results in the false prediction of an active KR domain in module 3 resulting in the prediction of a hydroxyl group rather than the correct keto group. The substrate choice of the loading domain was not predicted, but there was correct prediction of a C3 unit for five of the six extender modules; no prediction of substrate was possible for module 5. For niddamycin (Supplementary Data 2 Figure 2) there was also a false prediction of an additional ACP domain

in module 2. This results in a false positive prediction of an additional module and an inability to predict the chemical structures associated with the two 'modules'. In module 4, the KR domain was incorrectly predicted as active and the substrate for module 5 could not be predicted. Like *ClustScan* the wrong substrate for module 6 was predicted.

Eight further well-characterized clusters were annotated. For the megalomycin, pimaricin and tylactone clusters the predicted module activities were in full agreement with the published results. For tylactone (Supplementary Data 2 Figure 3) SEARCHPKS found all the domains, but is unable to predict activity of reduction domains; thus, it predicts a chemistry based on an active KR in module 4, whereas *ClustScan* correctly identifies the domain as inactive. Also, the starter unit is not correctly predicted. The worst results for *ClustScan* were obtained with the rifamycin cluster, where the stereochemistry of three methyl groups could not be predicted and two of eight DH domains were falsely predicted as active. In comparison, SEARCHPKS falsely predicts five DH domains and one KR domain as active and does not attempt to predict the stereochemistry (see Supplementary Data 2 Figure 4). For the other four clusters (amphotericin, avermectin, nystatin and oleandomycin) there were fewer errors (data not shown). Six additional domains were identified, which were not present in the published annotations. Two were TE domains; as the presence of a TE domain does not directly affect the structure of the compound, it is likely that previous annotation work had not searched carefully for these domains. The other four new domains were all DH domains with significant deletions (a third to a half of the length). They are, thus, predicted as inactive by *ClustScan*. Although such partially deleted domains are not important for prediction of product structure, they are interesting for studies on the evolution of clusters.

A major problem with annotations in DNA database entries is that they are not uniform, but differ according to the person carrying out the annotation. *ClustScan* helps achieve a uniform annotation standard and we have reannotated published sequences to achieve a standard definition of domain boundaries and description of units. *ClustScan* has been used to annotate successfully more than 50 modular gene clusters from a variety of genomes and metagenomes; full details are available on request.

The speed and convenience of *ClustScan* were assessed using the genome sequences of *Saccharopolyspora erythraea* (7) which is 8.2 Mb in size. A graduate student was able to annotate the PKS and NRPS clusters in 2–3 h of work (the initial analysis using HMMER can take several hours of run time on the server, but this occurs unsupervised overnight). The *ClustScan* annotation identified genes, modules and protein domains and included prediction of activity, substrate specificity and stereochemical outcome for PKS domains. The published annotation (7) identified genes, modules and protein domains and, in addition, the AT domains are assigned to malonyl–CoA and methylmalonyl–CoA-incorporating classes. However the stereochemistry and activity of reduction domains are not annotated. The *ClustScan* annotation agreed with the published annotation and extended it with predictions of domain activity and stereochemistry of products. *ClustScan* has been used to annotate DNA sequences from a variety of bacterial species including cyanobacteria.

ClustScan is mainly designed for use with bacterial sequences. However, the more general utility of *ClustScan* program package was demonstrated by the analysis of lower eukaryote sequences, where intron prediction is often difficult. An example is provided by the slime mould *Dictyostelium discoideum* which has 45 PKS genes (29), which were annotated poorly by the standard annotation methods used in the genome project. Using *ClustScan* it was possible to use local HMMER profiles for the protein domains, which are effective in recognizing segments of the domains split by introns. When such an analysis is carried out, a PKS gene shows a characteristic signature with parts of protein domains in the correct order 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.

ClustScan is mainly designed for the annotation of gene clusters encoding modular biosynthetic enzymes, but it can also be used for annotating other genes by loading appropriate HMMER profiles. For instance, we have used seven profiles to find and annotate shikimic acid pathway genes in a marine organism (30). Recently there has been intensive activity with metagenomic sequences. The source organisms for sequences are not known, but they contain genomes from a number of culturable and non-culturable microorganisms. The contigs are often fairly small and the quality of the sequence is sometimes poor. These problems make an analysis using HMMER local profiles attractive. We used *ClustScan* to analyze a 200 kb DNA sequence (AACY020563593) from the J. Craig Venter Institute Global Ocean Sampling (GOS) Expedition metagenomic dataset (31). This revealed a potential PKS–NRPS hybrid gene cluster of about 50 kb in size (Figure 4). It starts with an NRPS loading module, followed by three PKS modules and seven NRPS modules and ends with an NRPS thioesterase domain. However, closer examination of the domain distribution between reading frames reveals several cases where domains forming a single module appear to be present in different neighboring genes. This is due to three apparent frameshifts and the anomalous occurrence of a stop codon, which probably arise due to sequencing errors. Thus, it seems likely that there are three genes rather than the seven genes indicated by both GenMark and Glimmer analysis. In the case of two of the potential PKS modules, no AT domains are recognized, but there are unassigned regions in the protein of appropriate sizes and locations for AT domains (Figure 4). Thus, the program allows rapid scanning of metagenomic datasets and makes it easy to identify potential sequencing errors and interesting features of clusters. With the growing importance of metagenomic data for drug discovery programs *ClustScan* helps to eliminate a major bottleneck in the analysis.

6890 Nucleic Acids Research, 2008, Vol. 36, No. 21

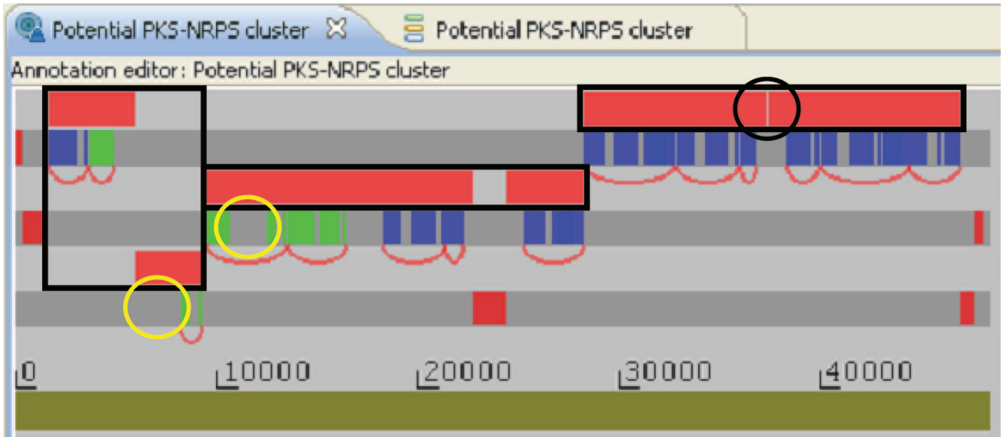

Figure 4. Annotation editor window showing the analysis of a potential PKS–NRPS hybrid cluster from a marine metagenomic sequence. The following coloring is used: genes (red), PKS protein domains (green) and NRPS protein domains (blue). Although seven genes are shown, the distribution of domains between genes suggest that sequencing errors have occurred. The three boxes indicate the positions of the probable genes. The first gene has one frameshift, the second gene has two frameshifts and the third gene has an anomalous stop codon (ringed in black) in it. The positions where two AT domains would be expected are also ringed (in yelow).

DISCUSSION

ClustScan is easy to use and allows rapid annotation of new gene clusters. This is very important for exploitation of the rapidly accumulating data from large-scale DNA sequencing projects. The facts that high-quality annotation with traditional methods is very time consuming and needs a high degree of experience have prevented full exploitation of the extensive DNA database to identify potentially interesting biosynthetic enzyme clusters. Although *ClustScan* is easy to use, it also allows the user to customize the result and override the automatic predictions. It is also designed to allow easy incorporation of new knowledge to improve predictive power. The server–client architecture means that such improvements as well as changes to reflect new versions of the standard analysis programs are implemented on the server and do not need changes in the client programs installed on users' computers. An important goal in the design of *ClustScan* was to give it an open architecture which would allow easy integration with other programs. The definition of an XML format for full gene cluster description allows interchange with other programs by simply adding an appropriate XML parser. The export of annotation as EMBL or GenBank formats and the export of chemical structures as SMILES (27) facilitate further analyses of results generated by *ClustScan*.

Knowledge about PKS protein domains is used to make predictions about chemical structure. In the case of the KR domain (14) there is detailed knowledge about protein structure and the role of the small number of amino acid residues that control reductase activity and stereospecificity. In the case of AT extender domains 13 amino acid residues that correlate with the choice between malonyl–CoA and methylmalonyl–CoA substrates were known (8–12). We found that these 13 amino acids could also be used to predict ethylmalonyl–CoA substrate. The incorporation of methoxymalonyl–CoA substrates was correlated with insertions. Initially, we tried to use a method similar to that of Minowa *et al.* (21) based on HMMER profiles of critical amino acids to predict AT specificity. However, this approach gave lower accuracy of prediction than the fingerprint method that we used subsequently. For both the KR and AT domains, the frequency of false prediction was low (4%). It was striking that good results were obtained for both Gram-negative sequences as well as for the majority of Gram-positive actinomycete sequences. This supports the idea that the diagnostic residues in AT domains are functional in substrate specificity rather than being evolutionary accidents. In contrast, the DH activity prediction, which was based on an actinomycete profile was only efficient for actinomycetes. In particular, many active Gram-negative DH domains were predicted to be inactive. This means that the profile mismatch is caused by the evolutionary distance. Although it would be possible in the short term to improve DH prediction using profiles for specific groups of organisms, the identification of important functional amino acid residues would give predictions less dependent on evolutionary distance. In contrast to other annotation programs (18,20,21,23,32), *ClustScan* predicts the stereochemistry of products. The dependence on functional residues in the KR and AT domains makes it especially valuable for novel gene clusters that are not closely related to known gene clusters. Such clusters are especially interesting in the search for novel drugs. We have not implemented specificity predictions for NRPS protein domains. However, there is some information available to allow partial prediction (32). When the prediction power is good enough it will be easy to add NRPS predictions to *ClustScan* and predict the chemical structure of products. We compared the performance of *ClustScan* to that of the SEARCHPKS prediction program of the NRPS–PKS database (18). This is less convenient to use as the genes must be identified and the deduced protein sequence input to the program. The output of predicted chemistry is not available in a standard chemical format. SEARCHPKS often predicts additional ACP domains

that prevent accurate prediction of product chemistry. We also observed that BLAST (19) searches often gave problems in identifying ACP domains, whereas no problems were encountered with HMMER (22); this is probably because of the short length of ACP domains. The prediction of the specificity of extender AT domains is relatively good; this probably reflects the fact that AT specificity correlates well with phylogenetic trees (33) so that, in addition to critical functional amino acids, there are other amino acids that differ for evolutionary reasons. As the BLAST program does not weight residues according to conservation, it works best when differences at many residues correlate with activity. SEARCHPKS does not give good prediction of loading module specificities. It does not attempt to predict activity or stereochemistry of domains. As these predictions involve a small number of critical residues, they could not be effectively implemented using a BLAST-based approach. The ASMPKS database (20) could not be meaningfully compared to *ClustScan* as it's gene prediction for clusters with high G + C-content was very poor and it requires a DNA input. This is because it uses the Glimmer (25) program to predict genes and builds an HMM-model from input data. In *ClustScan*, we implemented the use of custom HMM-models to overcome this difficulty for subgenomic sequences. As the ASMPKS implements a similar approach to the NRPS–PKS database, it is likely that similar results would be found if this technical problem were overcome.

There are at least 15 known starters used by different modular PKSs. In many cases there is no AT domain in the loading domain. Acetyl, propionyl and methylbutyryl starters can be loaded by AT domains and it was found that they could be distinguished using diagnostic amino acid residues. It was striking that the acetyl and propionyl starter AT domains showed the same patterns as the malonyl–CoA and methylmalonyl–CoA extender domains. It is known that in some cases an acetyl starter is derived from decarboxylation of a malonyl–CoA substrate, but in other cases acetyl–CoA is the substrate (34). The fact that the commonest extenders' AT domains are closely related to starter AT domains suggests that it might be possible to evolve new PKS gene clusters from truncated clusters that have lost the starter module.

Most polyketides undergo cyclization. In *ClustScan* we have implemented a simple rule of cyclization by interaction of the first hydroxyl or amino group with the terminal group. This applies to many natural polyketides and raises the hope that a simple rule-based method can make correct predictions in many cases. Prediction of cyclization is important to obtain the full benefit of product prediction.

The ability to rapidly acquire knowledge of new gene clusters from their DNA sequence has a variety of implications in the search for pharmacologically relevant compounds. The identification of novel gene clusters with interesting and unusual product chemistry will direct the choice of targets for lead discovery. Another application of the new sequences is to use them to construct new polyketides based on known modules *in silico*; i.e. use them as input for a program such as the *Biogenerator* program (35). *ClustScan* will help eliminate the bottleneck posed by the annotation of DNA sequences and allow the full utilization of the rapidly increasing DNA sequence data. Studies on the evolution of secondary metabolite clusters (36) can reveal biological constraints on the structures that can be attained; such studies are greatly assisted by the ability to rapidly and accurately annotate new clusters.

We used a top-down approach based on HMM models to annotate gene clusters encoding modular biosynthetic enzymes. We showed that by choice of appropriate profiles, *ClustScan* could also be used for annotating other primary and secondary metabolic pathways in a variety of microbial and invertebrate organisms. It seems likely that extensions of this approach could be useful for more general annotation tasks.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

ACKNOWLEDGEMENTS

We would like to thank Janko Diminic and Vedran Rodic for programming and Professor Arnold Demain for helpful advice and encouragement. J.S. declares a financial interest related to this work with regard to the potential future commercial marketing of the software and databases.

FUNDING

Ministry of Science, Education and Sports, Republic of Croatia (grant 058-0000000-3475 to D.H.); the German Academic Exchange Service (DAAD) and the Ministry of Science, Education and Sports, Republic of Croatia cooperation grant (to D.H. and J.C.); the Leverhulme Trust; Japanese Bio-Industry Association; The School of Pharmacy, University of London (to P.F.L.).

Conflict of interest statement. J.S. declares a financial interest related to this work with regard to the potential future commercial marketing of the software and databases.

REFERENCES

1. Challis,G.L. (2005) A widely distributed bacterial pathway for siderophore biosynthesis independent of nonribosomal peptide synthetases. *Chembiochem*, **6**, 601–611.
2. Finking,R. and Marahiel,M.A. (2004) Biosynthesis of nonribosomal peptides. *Ann. Rev. Microbiol.*, **58**, 453–488.
3. Hranueli,D., Cullum,J., Basrak,B., Goldstein,P. and Long,P.F. (2005) Plasticity of the *Streptomyces* genome - evolution and engineering of new antibiotics. *Curr. Med. Chem.*, **12**, 1697–1704.
4. Weissman,K.J. and Leadlay,P.F. (2005) Combinatorial biosynthesis of reduced polyketides. *Nat. Rev. Microbiol.*, **3**, 925–936.
5. Bentley,S.D., Chater,K.F., Cerdeno-Tarraga,A.M., Challis,G.L., Thomson,N.R., James,K.D., Harris,D.E., Quail,M.A., Kieser,H., Harper,D. *et al.* (2002) Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). *Nature*, **417**, 141–147.
6. Ikeda,H., Ishikawa,J., Hanamoto,A., Shinose,M., Kikuchi,H., Shiba,T., Sakaki,Y., Hattori,M. and Omura,S. (2003) Complete genome sequence and comparative analysis of the industrial microorganism *Streptomyces avermitilis*. *Nat. Biotechnol.*, **21**, 526–531.

7. Oliynyk,M., Samborskyy,M., Lester,J.B., Mironenko,T., Scott,N., Dickens,S., Haydock,S.F. and Leadlay,P.F. (2007) Complete genome sequence of the erythromycin-producing bacterium *Saccharopolyspora erythraea* NRRL23338. *Nat. Biotechnol.*, **25**, 447–453.
8. Haydock,S.F., Aparicio,J.F., Molnar,I., Schwecke,T., Khaw,L.E., König,A., Marsden,A.F., Galloway,I.S., Staunton,J. and Leadlay,P.F. (1995) Divergent sequence motifs correlated with the substrate specificity of (methyl)malonyl-CoA: acyl carrier protein transacylase domains in modular polyketide synthases. *FEBS Lett.*, **374**, 246–248.
9. Lau,J., Fu,H., Cane,D.E. and Khosla,C. (1999) Dissecting the role of acyltransferase domains of modular polyketide synthases in the choice and stereochemical fate of extender units. *Biochemistry*, **38**, 1643–1651.
10. Reeves,C.D., Murli,S., Ashley,G.W., Piagentini,M., Hutchinson,C.R. and McDaniel,R. (2001) Alteration of the substrate specificity of a modular polyketide synthase acyltransferase domain through site-specific mutations. *Biochemistry*, **25**, 15464–15470.
11. Del Vecchio,F., Petkovic,H., Kendrew,S.G., Low,L., Wilkinson,B., Lill,R., Cortes,J., Rudd,B.A.M., Staunton,J. and Leadlay,P.F. (2003) Active-site residue, domain and module swaps in modular polyketide synthases. *J. Ind. Microbiol. Biotechnol.*, **30**, 489–494.
12. Yadav,G., Gokhale,R.S. and Mohanty,D. (2003) Computational approach for prediction of domain organization and substrate specificity of modular polyketide synthases. *J. Mol. Biol.*, **328**, 335–363.
13. Caffrey,P. (2003) Conserved amino acid residues correlating with ketoreductase stereospecificity in modular polyketide synthases. *Chembiochem*, **4**, 654–657.
14. Reid,R., Piagentini,M., Rodriguez,E., Ashley,G., Viswanathan,N., Carney,J., Santi,D.V., Hutchinson,C.R. and McDaniel,R. (2003) A model of structure and catalysis for ketoreductase domains in modular polyketide synthases. *Biochemistry*, **42**, 72–79.
15. Starcevic,A., Cullum,J., Jaspars,M., Hranueli,D. and Long,P.F. (2007) Predicting the nature and timing of epimerisation on a modular polyketide synthase. *Chembiochem*, **8**, 28–31.
16. Castonguay,R., He,W., Chen,A.Y., Khosla,C. and Cane,D.E. (2007) Stereospecificity of ketoreductase domains of the 6-deoxyerythronolide B synthase. *J. Am. Chem. Soc.*, **129**, 13758–13769.
17. Keatinge-Clay,A.T. (2007) A tylosin ketoreductase reveals how chirality is determined in polyketides. *Chem. Biol.*, **14**, 898–908.
18. Yadav,G., Gokhale,R.S. and Mohanty,D. (2003) SEARCHPKS: a program for detection and analysis of polyketide synthase domains. *Nucleic Acids Res.*, **31**, 3654–3658.
19. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
20. Tae,H., Kong,E.B. and Park,K. (2007) ASMPKS: an analysis system for modular polyketide synthases. *BMC Bioinformatics*, **8**, 327.
21. Minowa,Y., Araki,M. and Kanehisa,M. (2007) Comprehensive analysis of distinctive polyketide and nonribosomal peptide structural motifs encoded in microbial genomes. *J. Mol. Biol.*, **368**, 1500–1517.
22. Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
23. Zazopoulos,E., Huang,K., Staffa,A., Liu,W., Bachmann,B.O., Nonaka,K., Ahlert,J., Thorson,J.S., Shen,B. and Farnet,C.M. (2003) A genomics-guided approach for discovering and expressing cryptic metabolic pathways. *Nat. Biotechnol.*, **21**, 187–190.
24. Besemer, J. and Borodovsky, M. (2005) GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. *Nucleic Acids Res.*, **33**, Web Server issue W451–454.
25. Delcher,A.L., Bratke,K.A., Powers,E.C. and Salzberg,S.L. (2007) Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics*, **23**, 673–679.
26. Bateman,A., Birney,E., Cerruti,L., Durbin,R., Etwiller,L., Eddy,S.R., Griffiths-Jones,S., Howe,K.L., Marshall,M. and Sonnhammer,E.L. (2002) The Pfam protein families database. *Nucleic Acids Res.*, **30**, 276–280.
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.
28. Haydock,S.F., Appleyard,A.N., Mironenko,T., Lester,J., Scott,N. and Leadlay,P.F. (2005) Organization of the biosynthetic gene cluster for the macrolide concanamycin A in *Streptomyces neyagawaensis* ATCC 27449. *Microbiology*, **151**, 3161–3169.
29. Zucko,J., Skunca,N., Curk,T., Zupan,B., Long,P.F., Cullum,J., Kessin,R. and Hranueli,D. (2007) Polyketide synthase genes and the natural products potential of *Dictyostelium discoideum*. *Bioinformatics*, **23**, 2543–2549.
30. Starcevic,A., Akthar,S., Dunlap,W.C., Shick,J.M., Hranueli,D., Cullum,J. and Long,P.F. (2008) 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**, 2533–2537.
31. Rusch,D.B., Halpern,A.L., Sutton,G., Heidelberg,K.B., Williamson,S., Yooseph,S., Wu,D., Eisen,J.A., Hoffman,J.M., Remington,K. *et al.* (2007) The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through Eastern Tropical Pacific. *PLoS Biol.*, **5**, e77.
32. Rausch,C., Weber,T., Kohlbacher,O., Wohlleben,W. and Huson,D.H. (2005) Specificity prediction of adenylation domains in nonribosomal peptide synthetases (NRPS) using transductive support vector machines (TSVMs). *Nucleic Acids Res.*, **33**, 5799–5808.
33. Jenke-Kodama,H., Börner,T. and Dittmann,E. (2006) Natural biocombinatorics in the polyketide synthase genes of the actinobacterium *Streptomyces avermitilis*. *PLoS Comput. Biol.*, **2**, e132.
34. Long,P.F., Wilkinson,C.J., Bisang,C.P., Cortes,J., Dunster,N., Oliynyk,M., McCormick,E., McArthur,H., Mendez,C., Salas,J.A. *et al.* (2002) Engineering specificity of starter unit selection by the erythromycin-producing polyketide synthase. *Mol. Microbiol.*, **43**, 1215–1225.
35. Zotchev,S.B., Stepanchikova,A.V., Sergeyko,A.P., Sobolev,B.N., Filimonov,D.A. and Poroikov,V.V. (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.
36. Fischbach,M.A., Walsh,C.T. and Clardy,J. (2008) The evolution of gene collectives: How natural selection drives chemical innovation. *Proc. Natl Acad. Sci. USA*, **105**, 4601–4608.

2.3 Proposed arrangement of proteins forming a bacterial type II polyketide synthase

Gaetano Castaldo*, **Jurica Zucko***, Sibylle Heidelberger, Dušica Vujaklija, Daslav Hranueli, John Cullum, Pakorn Wattana-Amorn, Matthew P. Crump, John Crosby and Paul F. Long. *Chem. Biol.*, ***15***, 1156-1165, 2008.

Abstract:

Aklanonic acid is the first enzyme free intermediate in the biosynthesis of daunorubicin and doxorubicin, two of the most widely used anticancer agents. Aklanonic acid is synthesised on a type II polyketide synthase (PKS) composed of 8 proteins that catalyze the condensation between a propionate starter unit and 9 malonate extender units. This type II PKS is unique in that it contains a type III ketosynthase (DpsC) that chooses the starter unit and a putative malonyl/acetyl transferase (DpsD) whose role seems obscure. We have investigated the network of protein interactions within this complex using a yeast two hybrid system, co-affinity chromatography and by computer aided protein docking simulations. Our results suggest that the ketosynthase (KS) α and β subunits interact with each other and that the KSα subunit (DpsA) also probably interacts with DpsD forming a putative minimal synthase. We speculate that DpsD may physically inhibit the priming reaction allowing the choice of propionate rather than acetate as the starter unit. We also suggest a structural role for the cyclase (DpsY), perhaps maintaining the overall structural integrity of the complex. This represents the first study attempting to analyze *in vivo* protein interactions forming a type II PKS.

Own contribution to the paper:

Homology modelling and rigid protein-protein docking simulations of enzymes involved in biosynthesis of polyketide daunorubicin.

* These authors contributed equally to this work

Chemistry & Biology
Article

Proposed Arrangement of Proteins Forming a Bacterial Type II Polyketide Synthase

Gaetano Castaldo,[1,6] Jurica Zucko,[2,3,6] Sibylle Heidelberger,[1] Dušica Vujaklija,[4] Daslav Hranueli,[2] John Cullum,[3] Pakorn Wattana-Amorn,[5] Matthew P. Crump,[5] John Crosby,[5] and Paul F. Long[1,*]
[1]School of Pharmacy, University of London, 29-39 Brunswick Square, Bloomsbury, London WC1N 1AX, UK
[2]Faculty of Food Technology and Biotechnology, University of Zagreb, Pierottijeva 6, 10000 Zagreb, Croatia
[3]Department of Genetics, University of Kaiserslautern, Postfach 3049, 67653 Kaiserslautern, Germany
[4]Department of Molecular Biology, Ruđer Bošković Institute, PO Box 180, 10002 Zagreb, Croatia
[5]School of Chemistry, University of Bristol, Cantock's Close, Bristol BS8 1TS, UK
[6]These authors contributed equally to this work
*Correspondence: paul.long@pharmacy.ac.uk
DOI 10.1016/j.chembiol.2008.09.010

SUMMARY

Aklanonic acid is synthesized by a type II polyketide synthase (PKS) composed of eight protein subunits. The network of protein interactions within this complex was investigated using a yeast two-hybrid system, by coaffinity chromatography and by two different computer-aided protein docking simulations. Results suggest that the ketosynthase (KS) α and β subunits interact with each other, and that the KSα subunit also probably interacts with a malonyl-CoA: ACP acyltransferase (DpsD), forming a putative minimal synthase. We speculate that DpsD may physically inhibit the priming reaction, allowing the choice of propionate rather than acetate as the starter unit. We also suggest a structural role for the cyclase (DpsY) in maintaining the overall structural integrity of the complex.

INTRODUCTION

The diversity of biological processes is due to dynamic associations between cellular components, including noncovalent protein-protein and protein-ligand interactions (Parrish et al., 2006). For a number of metabolic pathways, several enzymes that catalyze sequential reactions often associate noncovalently to form a multienzyme complex. Such complexes afford increased reaction rates and protect labile intermediates from decomposition by channeling intermediates directly from one active site to another. Studies on enzyme complex formation and substrate channeling are essential for a better understanding of metabolism. To achieve this, we need to know the reactive groups involved at the active sites of the enzymes, the amino acids involved in surface binding sites, the specific order in which the large protein complexes are assembled, and the overall topology of the complex. Polyketides are a large and structurally diverse group of natural products that display an impressive range of biological activities of major economic importance to the pharmaceutical and agrochemical industries. These compounds are synthesized by large multienzyme systems called polyketide synthases (PKSs) that catalyze the sequential decarboxylative condensation between short chain coenzyme A (CoA)-derived carboxylic acids by a mechanism analogous to fatty acid biosynthesis. The growing carbon chain backbone then undergoes regio- and stereoselective modification to give the final natural product.

Type II PKSs consist of several discrete, monofunctional proteins that form a dissociable complex, usually leading to the biosynthesis of aromatic polyketides (Hertweck et al., 2007). A minimal type II PKS is formed on association of a ketosynthase (KS), termed KSα, a KS homolog lacking the active site cysteine, often referred to as the chain length factor (CLF) or KSβ, and an acyl carrier protein (ACP). This minimal system controls starter unit selection, chain length, and the first cyclization of the nascent polyketide chain. A fourth enzyme, malonyl-CoA:ACP acyltransferase (MCAT), may be required for substrate loading in vivo (Revill et al., 1995; Summers et al., 1995), although the demonstration of an inherent malonyl transferase activity of the type II ACP indicates that MCAT may not be needed in vitro (Hitchman et al., 1998; Matharu et al., 1998). Additional enzymes, such as ketoreductase (KR), cyclase (CYC), and aromatase (ARO), associate with the minimal complex to generate aromatic natural products (McDaniel et al., 1995; Kramer et al., 1997; Funa et al., 1999; Petkovic et al., 1999). There is currently very little detailed information about the three-dimensional (3D) organization of type II PKS complexes, a factor that undoubtedly limits the rational design of novel polyketides in these systems. Much more information is available on individual protein structures associated with type II PKSs. The X-ray structure of the heterodimeric KS/CLF from *Streptomyces coelicolor* has been solved and the cavity that determines chain length identified (Keatinge-Clay et al., 2003). A number of solution structures for PKS ACPs are available (Crump et al., 1997; Findlow et al., 2003; Li et al., 2003), although, as yet, the exact nature of the protein-protein interactions between the carrier protein and the KS/CLF heterodimer that allow the formation of an active minimal complex remain to be identified. Several auxiliary enzymes that may interact either with the minimal complex or the ACP component of the complex have also been structurally characterized. These include the KR from *S. coelicolor* (Hadfield et al., 2004; Korman et al., 2004), the methyltransferase from *Streptomyces peucetius* (Jansson et al., 2004), and the CYC from *Streptomyces nogalater* (Sultana et al., 2004), *Streptomyces glaucescens* (Thompson et al., 2004), and *Streptomyces galilaeus* (Sultana et al., 2004). The CYC may be particularly important for

Chemistry & Biology
Type II Polyketide Synthase Structural Studies

Table 1. Matrix of Possible Protein-Protein Interactions for the Entire Daunorubicin/Doxorubicin-Producing PKS Measured Using a Y2H Assay

	Bait							
Prey	A (KSα)	B (KSβ)	C (KSIII)	D (MAT)	G (ACP)	E (KR)	F (ARO)	Y (CYC)
A (KSα)	AA++	AB++	AC	AD	AG+	AE	AF+	AY+
B (KSβ)	BA	BB	BC	BD+	BG+	BE+	BF+	BY
C (KSIII)	CA	CB	CC	CD	CG+	CE+	CF+	CY+
D (MAT)	DA++	DB+	DC	DD	DG+	DE++	DF+	DY+
G (ACP)	GA	GB	GC	GD	GG	GE+	GF	GY+
E (KR)	EA	EB	EC	ED+	EG+	EE	EF+	EY+
F (ARO)	FA+	FB	FC	FD+	FG+	FE+	FF+	FY+
Y (CYC)	YA++	YB++	YC+	YD++	YG	YE++	YF	YY++

ACP, acyl carrier protein; ARO, aromatase; CYC, cyclase; KR, ketoreductase; KS, ketosynthase; MCAT, malonyl-CoA:ACP acyltransferase. "++" indicates strong interactions revealed by nutritional selection, while "+" indicates weak interactions revealed by *LacZ* β-galactosidase assay.

the stability of the overall complex, as its addition eliminates the production of shorter polyketides as well as increasing the turnover of the complex (Yu et al., 1998).

Daunorubicin (DNR) and the C-14 hydroxylated derivative, doxorubicin, are among the most widely used antitumor anthracyclines (Minotti et al., 2004). Both anthracyclines are produced by *S. peucetius* through a pathway involving a type II PKS (Keatinge-Clay et al., 2004). This PKS catalyzes the condensation between a propionyl-CoA starter unit and nine malonyl-CoA extender units, producing a 21 carbon decaketide. Aldol condensation followed by C-12 oxidation of the decaketide leads to the formation of the first enzyme-free intermediate, aklanonic acid. The gene cluster encoding DNR biosynthesis consists of eight genes, designated *dpsA*, *-B*, *-C*, *-D*, *-G*, *-E*, *-F*, and *-Y* (Grimm et al., 1994). Genes *dpsA* and *-B* encode the KS and CLF enzymes, while the ACP is encoded by *dpsG*, unusually positioned 6.8 kb upstream of the position seen in other type II PKSs. This PKS is also unusual in the choice of a propionate rather than an acetate starter unit (Rajgarhia and Strohl, 1997), with the enzymes encoded by *dpsC* and *-D* playing a crucial role in the specification of this starter unit (Bao et al., 1999a, 1999b). The enzyme encoded by *dpsC* is a homolog of the β-ketoacyl: ACP synthase III (KASIII) responsible for the condensation between the starter unit and the first extender unit, while *dpsD* encodes a proposed MCAT (Rajgarhia and Strohl, 1997). The genes *dpsC* and *-D* are rare, and equivalent enzymes have only been described in this and other type II PKS clusters that utilize nonacetate starters (Bibb et al., 1994; Piel et al., 2000; Raty et al., 2002). Their role in starter unit selection is not, however, entirely clear, as deleting *dpsC* but not *dpsD* shifted starter unit selection from propionate to predominantly acetate (Rajgarhia and Strohl, 1997), suggesting that *dpsC* and not *dpsD* contributes to, but does not dictate, starter unit selection. The *dpsE* gene product is a KR, with *dpsF* coding for an ARO (Meurer et al., 1997). Although initial studies failed to identify the function of *dpsY* (Lomovskaya et al., 1998), it was known to be essential for the production of DNR in *S. peucetius*, as its deletion leads to the formation of aberrant cyclization products. Its function as a CYC was later confirmed (Wohlert et al., 2001). To date, few of the type II DNR PKS enzymes have been expressed and purified; none have been structurally characterized. In this article, we extend our initial studies (Castaldo et al., 2005) to obtain further information on the in vivo protein-protein interactions involved in the biosynthesis of aklanonic acid.

RESULTS AND DISCUSSION

Investigating Protein-Protein Interactions Using a Yeast Two-Hybrid System

To investigate interactions between proteins forming the minimal aklanonic acid-producing PKS (Grimm et al., 1994; Hutchinson and Colombo, 1999), the genes encoding the KS subunits (α, *dpsA*; β, *dpsB*), the ACP (*dpsG*), the KASIII (*dpsC*), and MCAT (*dpsD*) were cloned and assayed as both prey and bait using a matrix of all possible protein interactions (Table 1). All potential interactions were tested independently three times along with the relative controls. Strong interactions were assessed by nutritional selection using *ADE2* and *HIS3* markers. The results from these assays (Table 1) suggest that DpsA (KSα) interacted with DpsB (KSβ or CLF). Such an interaction between the KS subunits has been described for many complexes. It has also been known for some time that, if the equivalent subunits from other type II PKS systems are expressed and purified, they coelute, suggesting strong, noncovalent interactions (Carreras and Khosla, 1998; Matharu et al., 1998). These interactions can be identified from the crystal structure of the actinorhodin KS/CLF heterodimer (Keatinge-Clay et al., 2004). The assays also suggest that KSα (DpsA) interacts strongly with itself, in contrast to the actinorhodin KSα, which is monomeric (Keatinge-Clay et al., 2004). Structural studies of the actinorhodin KSα-KSβ have demonstrated that these two proteins form an amphipathic tunnel, with polyketide synthesis at the heterodimer interface (Keatinge-Clay et al., 2004). This structure is thought to predict overall chain length as well as partially specifying correct first-ring cyclization. The aklanonic acid-producing KSα-KSβ proteins also show strong interactions and, by analogy with the actinorhodin system, may dictate chain length and correct cyclization in a similar fashion. It has been suggested that the actinorhodin KSβ acts as a malonyl-CoA decarboxylase, thereby generating the acetyl-ACP starter for this biosynthetic pathway (Bisang et al., 1999). This has not been unanimously accepted, with the KSα implicated as the alternative source of the decarboxylase activity (Dreier and Khosla, 2000), nor has such an activity been identified in KSβ (DpsB). It is known that acetyl-CoA can act as a starter unit in

anthracycline biosynthesis, suggesting that decarboxylation by either of the KS subunits would be unnecessary (Rajgarhia and Strohl, 1997). Decarboxylation has not been ruled out, however, and the weak interactions between ACP and both of the KS subunits suggest that this may still be a possibility.

Interestingly, the nutritional selection assays also revealed a strong interaction between KSα (DpsA) and the MCAT, DpsD, but its role in aklanonic acid biosynthesis is not clear, as deletion of *dpsD* from the gene set does not affect compound production (Grimm et al., 1994). In addition, initial experiments involving heterologous expression of all of the genes involved in aklanonic acid production, but excluding *dpsC* and *dpsD*, seemed to result in synthesis of the correct tricyclic 21 carbon intermediate (Rajgarhia and Strohl, 1997). However, the polyketide SEK43 was produced in subsequent in vitro studies, indicating the use of acetyl, and not propionyl-CoA, to form this aberrant cyclized product (Meurer et al., 1997). The anthracycline, feudomycin D, could also be isolated from the *dpsC* and *dpsD* mutant (Rajgarhia and Strohl, 1997). This anthracycline is formed via desmethylaklanonic acid, again an acetate-initiated polyketide, with miscyclization to SEK43 prevented by the presence of additional cyclases. DNR is also produced in this system, but only at 40% of control levels. The correct propionate starter could only be ensured if *dpsC* (though not necessarily in tandem with *dpsD*) was present, confirming the suggestion that the *dpsC* gene product specifies propionyl starter unit selection (Grimm et al., 1994; Rajgarhia et al., 2001). It has been suggested that KASIII (DpsC), along with KSα (DpsA) and KSβ (DpsB), are together responsible for this process, so it is surprising that KASIII (DpsC) does not appear to interact strongly with the KS subunits, nor does it interact with MCAT (DpsD), suggesting that these are not a tandem pair of enzymes. The MCAT (DpsD) does not appear to select the CoA-derived polyketide starter, and its absence in vivo still leads to the production of aklanonic acid. In vitro, however, extracts containing all of the PKS genes, except MCAT (DpsD), failed to produce any polyketide products irrespective of whether propionyl- or acetyl-CoA were provided to initiate the biosynthesis (Rajgarhia et al., 2001), suggesting a structural role within the complex. This has also been observed for other type II systems that contain an MCAT (DpsD) homolog (Tang et al., 2004), suggesting instability in the PKS complex that may be compensated for in vivo by other cellular components. Indeed, crosstalk between the *S. coelicolor* FAS malonyl transferase and the actinorhodin PKS complex has previously been suggested (Summers et al., 1995; Raty et al., 2002). Alternatively, MCAT (DpsD) may act as an acyl-ACP thioesterase, which selectively hydrolyzes acetyl groups, thereby favoring propionyl starter selection. This activity has been described for ZhuC, a homologous enyzyme to MCAT (DpsD), which acts as part of the initiation module from the R1128-producing PKS (Tang et al., 2003).

No strong interactions between the *dpsG* gene product (the ACP), and either the KS subunits (DpsA and B), or KASIII (DpsC) or MCAT (DpsD), could be detected, suggesting that interactions between the ACP and these components of the complex are weak or transient. No phosphopantetheinyl (PPT) transferase (PPTase) has yet been identified that is involved in secondary metabolism in *Saccharomyces cerevisiae* (Wattanachaisaereekul et al., 2008). It has been demonstrated that heterologous expression of the 6-methylsalicylic acid synthase from *Penicillium patulum* in *S. cerevisiae* does require coexpression of an exogenous PPTase to convert apo-ACP to its holo form (Kealey et al., 1998; Wattanachaisaereekul et al., 2008). PPT phosphate may well be a major binding-energy contributor, so it is feasible that no strong interactions between the ACP (DpsG) and the other domains were observed because the ACP (DpsG) was in the apo form when expressed in *S. cerevisiae* for the yeast two-hybrid (Y2H) assay. For this reason, and in order to investigate potentially weaker interactions between the proteins of the minimal DNR PKS, *LacZ* assays were performed on combinations where no interactions could be observed by nutritional selection (shown in Table 1). When no interactions were detected either by nutritional selection or by *lacZ* assay, Western blots confirmed protein expression in the yeast heterologous host (data not shown). Using the *lacZ* assay, weak interactions were observed between the ACP (DpsG) and both of the KS subunits, as well as the KASIII homolog (DpsC) and the MCAT (DpsD). A homodimeric interaction between two ACPs was not observed. Such an interaction has been described for several type II PKS ACPs (Hitchman et al., 1998; Matharu et al., 1998; Florova et al., 2002), an interaction that facilitates the inter-ACP transfer of malonate. It is possible that, in the aklanonic acid-producing PKS, this acyl transfer is performed by another component of the complex, possibly MCAT (DpsD). ACPs are known to be essential for polyketide production in a number of type II minimal systems (McDaniel et al., 1995; Matharu et al., 1998), and it has been shown that the levels of ACP may be a limiting factor in the production of these secondary metabolites (Decker et al., 1994; Matharu et al., 1998). A model for the *S. coelicolor* actinorhodin minimal PKS complex has been described where the ACP dissociates from the KS/CLF after each round of condensation (Dreier and Khosla, 2000). This model would also be consistent with the observation that the ACP interacts weakly with the other components of the minimal complex. The weak interaction between the ACP and KASIII (DpsC) supports the hypothesis that the KASIII homolog acts in the priming reaction catalyzing the condensation of the starter unit, propionyl-CoA, to a malonyl-CoA extender unit with the product transferred to the 4′-phosphopantetheine thiol of the ACP. No homodimeric interactions involving KASIII (DpsC) were observed, as supported by previous studies (Bao et al., 1999a, 1999b), although this appears to be a unique feature in the aklanonic acid-producing PKS, since other KASIII-like proteins—for example, the FabH of *Escherichia coli* (Qin et al., 2001; Qiu et al., 2001) and ZhuH in R1128 biosynthesis (Pan et al., 2002)—show a homodimeric structure.

Finally, the proteins involved in auxiliary processing of the growing polyketide carbon chain, the KR (DpsE), the ARO (DpsF), and the CYC (DpsY) were investigated. As before, initial screening was performed by nutritional selection to indicate the strongest interactions, while weaker interactions were highlighted using the *LacZ* assay. Results, shown in Table 1, revealed that the CYC (DpsY) interacts strongly with the two subunits of the KS (DpsA and DpsB), with the MCAT (DpsD), and with the KR (DpsE). The CYC also interacts strongly with itself, suggesting either a dimeric or tetrameric arrangement, quaternary structures that have also been observed for the tetracenomycin F2 CYC from *S. glaucescens* (Thompson et al., 2004) and for SnoaL, the enzyme catalyzing the last cyclization step in *S. nogalater* (Beinker et al., 2006). A stronger interaction was also observed

Chemistry & Biology
Type II Polyketide Synthase Structural Studies

between MCAT (DpsD) and the KR (DpsE). For DpsE, there appears to be no indication of a dimeric form, which has been described for the actinorhodin KR (Hadfield et al., 2004; Korman et al., 2004). Weak interactions between all of the enzymes tested were observed, with the exception of KSα (DpsA) with KR (DpsE), and KR (DpsE) with itself.

As with all techniques, the interpretation of the results must take into account the limitations of the methods used, and the Y2H approach has a number of well-documented disadvantages. False-positive interactions are possible through autoactivation of the bait fusion. False-negative interactions may also arise through incorrect folding of either the bait or prey chimeric proteins. Many proteins also require posttranslational modifications in order to attain the correct structure and full biological activity, and the Y2H assay may fail to detect proteins whose interactions depend on such modifications. The necessity for phosophopantetheinylation of the ACP in order to generate the active form is well documented (Mootz et al., 2001), although covalent modifications have also been identified for other type II PKS enzymes. Some are subjected to proteolytic processing, while others show a combination of truncation and covalent addition (Gramajo et al., 1991; Hesketh et al., 2002). The Y2H system may also be unsuitable for the detection of interactions with membrane proteins, which may be improperly folded due to exposed, highly hydrophobic patches. This may be a particular problem for the KS, as this enzyme, which is central to the PKS minimal complex, may be membrane associated (Gramajo et al., 1991).

Investigating Protein-Protein Interactions Using Tandem Affinity Purification

From the Y2H results, KSα (DpsA) appeared to be central to the formation of a "minimal" PKS, which we speculate to be a homodimer composed of a head-to-tail arrangement of KSα (DpsA), KSβ (DpsB), and MCAT (DpsD). Tandem affinity purification (TAP) (Rigaut et al., 1999) was used to investigate this association further, with the TAP tag fused at the N terminus of KSα (DpsA). This technique allows purification of protein complexes under native conditions by using two different affinity purification steps. Expression of the minimal PKS with the hybrid protein TAP tag-DpsA was performed under the control of the strong constitutive promoter *ermE*p* (Carreras and Khosla, 1998) using the heterologous host *S. coelicolor* A3(2). Transcription of *dpsA*, *dpsB*, *dpsC*, *dpsD*, and *dpsG* were detected by RT-PCR. The presence of hybrid protein TAP tag-DpsA was detected by immunoblotting using IgG antibody that binds the ProtA epitope located at the N terminus of the TAP tag. Subsequent analysis of the protein eluates collected at the end of the purification step by SDS-PAGE revealed the presence of only KSα (DpsA) and KSβ (DpsB), which was confirmed by mass spectrometry. Proteins corresponding to KASIII (DpsC), MCAT (DpsD), or the ACP (DpsG) could not be detected by SDS-PAGE followed by staining with Coomassie brilliant blue (a one-dimensional [1D] SDS-PAGE gel is shown in the Supplemental Data available online—see Figure S1).

Strong interaction between KSα (DpsA) and KSβ (DpsB) was not unexpected, since similar interactions had been predicted from the X-ray crystal structure of KSα/KSβ from the actinorhodin-producing PKS (Keatinge-Clay et al., 2004) and by copurification of these proteins by gel chromatography as an $\alpha_2\beta_2$ heterotetramer (Carreras and Khosla, 1998). Comparison with the crystal structure of the actinorhodin KSα/KSβ complex would suggest that the N terminus of KSα (DpsA) is sufficiently exposed and not involved in crucial interactions with KSβ (DpsB). However, the presence of the TAP tag in this region of the protein might have impaired the interaction with the other "minimal" components, such as MCAT (DpsD). Failure to recover KASIII (DpsC) and the ACP (DpsG) might have been expected, since these proteins were found to form only weak interactions with the minimal PKS by the *lacZ* assay in the Y2H screen. This has also been suggested by the proposed mechanism of action of KASIII in type II fatty acid biosynthesis (Jackowski et al., 1989) and the ACP in the biosynthesis of thiolactomycin by *E. coli* (White et al., 2005).

Investigating Protein-Protein Interactions by Computer Simulation

ClusPro (Comeau et al., 2003) is a fully automated, Web-based program for docking protein structures. It is designed as a multistage protocol, which first performs rigid body searches using ZDOCK (Chen et al., 2003). ZDOCK uses fast Fourier transform to search all possible binding modes for the proteins. Its scoring functions combine shape complementarity, desolvation energy, and electrostatics in its calculations. Docked structures are then filtered using distance-dependent electrostatics and an empirical potential representing desolvation. The 2000 conformations retained after filtering are clustered based on pairwise root-mean-square deviation (rmsd), which is the measure of the average distance between the backbones of the superimposed proteins. The representative conformations from the 30 largest clusters are selected and refined using a brief CHARMm minimization (CHARMm is a program within ClusPro for macromolecular energy, minimization, and dynamics calculations). In our docking simulations, the first 10 cluster representatives were retained.

As second docking simulations, PatchDock (Schneidman-Duhovny et al., 2005), in conjunction with FireDock (Mashiach et al., 2008), were used to evaluate the results obtained by ZDOCK. PatchDock (Duhovny et al., 2002) is a geometry-based molecular docking algorithm. The PatchDock algorithm divides the Connolly dot surface representation of the molecules into concave, convex, and flat patches. Complementary patches are then matched in order to generate candidate transformations. Each candidate transformation is further evaluated by a scoring function that considers both geometric fit and atomic desolvation. FireDock (Andrusier et al., 2007) is a method for the refinement and rescoring of the rigid-body docking solutions. Each candidate generated by the rigid-body docking method was refined using a restricted interface side chain rearrangement and by soft, rigid-body optimization. Refined candidates are then ranked by the binding score, which includes atomic contact energy, softened van der Waals interactions, partial electrostatics, and additional estimations of the binding free energy. The output is a ranked list of all the input solutions. For docking simulations, PatchDock was used with default settings, and the first 100 solutions were refined using Firedock. The top 10 results from FireDock were retained.

The use of docking algorithms to investigate protein interactions requires knowledge of the tertiary structure of the putative

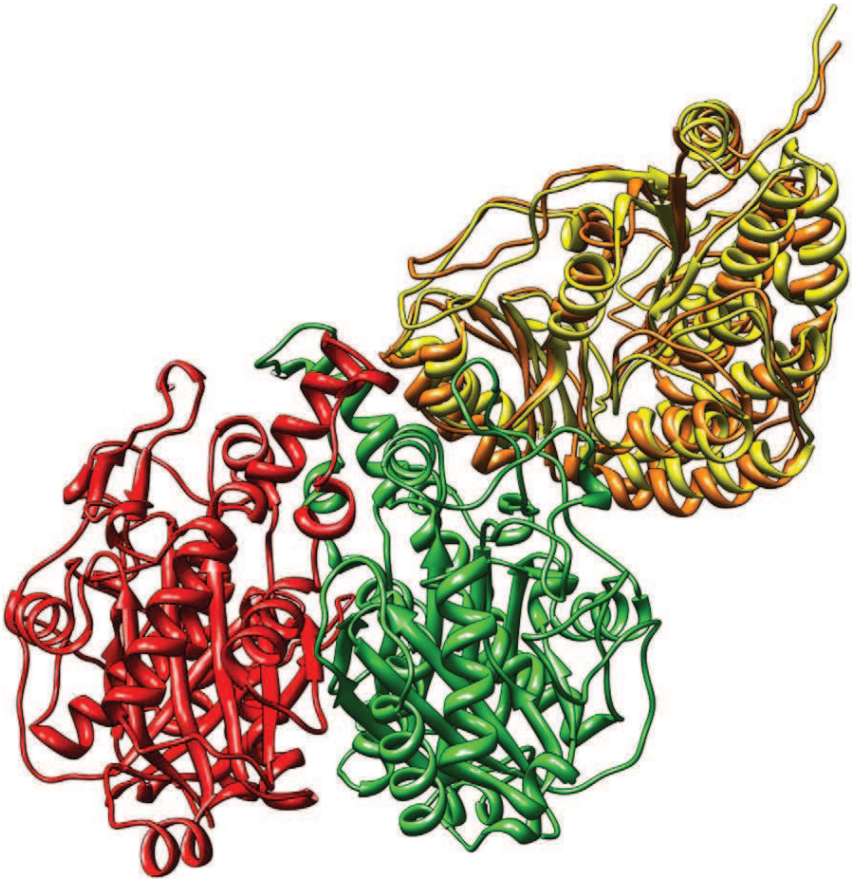

Figure 1. Computational Docking Simulating the Protein-Protein Interactions between DpsAB and DpsD

The DpsA (green)/DpsB (red) interaction matches the predicted 2.0 Å X-ray crystal structure of the KS subunits from the actinorhodin-producing PKS. The simulation also supports the Y2H (Table 1) results where a strong interactions between DpsA/DpsB and DpsA/DpsD were observed. These results also suggest that the failure to pull down an intact DpsAB/DpsD complex could be due to the design of the coaffinity chromatography, as discussed in the main text. DpsD is shown in yellow.

interacting proteins. However, the number of structurally characterized proteins is extremely low compared with the annotated primary structure of proteins present in databases (Zuiderweg, 2002; Bairoch et al., 2005). Comparative modeling is widely recognized as a reliable method to generate a 3D model of a target protein from its primary structure (Tramontano and Morea, 2003; Moult, 2005). An essential requirement for this method is the identification of at least one experimentally solved 3D structure of a protein homolog that can be used as the template.

Solutions could only be reasonably computed for interactions between DpsA/DpsB, DpsA/DpsD, DpsAB/DpsD, DpsAB/DpsY, DpsD/DpsY, and DpsD/DpsE. An interaction between KSα (DpsA), KSβ (DpsB), and the ACP (DpsG) could not be predicted using a monomeric model of KSα (DpsA) and KSβ (DpsB). Identifying the residues involved in the crucial interaction between the ACP and KS subunits will be a matter of important future research that will require more sophisticated biophysical methods (e.g., electron microscopy). Both programs returned solutions for KSα (DpsA) and KSβ (DpsB) models that matched the predicted 2.0 Å X-ray crystal structure of the KS subunits from the actinorhodin-producing PKS (Figure 1). It was the first solution from FireDock and the second from ClusPro. The Cα rmsd of superimposed solutions was 0.42 Å. Stroud and coworkers (Keatinge-Clay et al., 2003) reported that the two KS subunits interact via tight complementary contacts that bury over one-fifth of the surface area of each monomer, forming an amphipathic tunnel. The cavity at the interface of the two monomers is where the polyketide backbone is synthesized, and its 17.0 Å length influences the chain length of the growing polyketide (Keatinge-Clay et al., 2004). A "grasping loop" structure formed between the α7 helix of the KSα and α8 helix of the KSβ is responsible for the tight interactions between the two subunits. In particular, Tyr118 of KSα and the Phe116 of KSβ, were found to be involved in establishing close interactions. The active sites of the KSα, Cys169 and Phe116 are thought to represent the gating residues that regulate chain length marking the beginning and the end of the amphipathic tunnel, respectively (Keatinge-Clay et al., 2004). Tyr118 is conserved in the same position in the amino acid sequence of KSα (DpsA), whereas Phe116 of KSβ is substituted by leucine on position 118 (Leu118) in the model (position Leu138 in the primary sequence) of KSβ (DpsB). With the substitution of Phe with Leu, the length of the amphiphatic tunnel is increased to ~19 Å, possibly reflecting the difference in acyl chain length between actinorhodin and daunorubicin polyketides. In a similar fashion to the actinorhodin system, we can speculate, based on the docking simulation using the predicted 3D structures of KSα (DpsA) and KSβ (DpsB), that the aklanonic acid backbone is also synthesized in a polyketide tunnel, the length of which is determined by the distance between the two residues, Cys169 on DpsA and Leu118 on DpsB (Phe116 on act CLF). The docking simulation with two DpsA monomers revealed a complex that was comparable with the crystal structure of the actinohrodin KS-CLF complex and predicted DpsA-DpsB complex. This was the first solution from both docking methods. The Cα rmsd of superimposed solutions was 1.92 Å.

The proposed orientation (Figure 1) for the interaction between DpsAB/DpsD showed that docking of DpsD occurs in a pocket on DpsA created between helices α2 Pro57-Ala60, α3 Ala64-Arg69, α6 Thr111-Ser122, and the loop formed from residues Ser38-Arg46, which is located between helices α1 and α2. In this interaction, DpsD is involved with two α helices from small subunit α8 (Gly149–Asn15) and α9 (Val174–Leu184) and the N terminus of helix α10 (Pro203-Thr219), from which Pro203 and Met204 are involved in establishing hydrophobic interactions

Chemistry & Biology
Type II Polyketide Synthase Structural Studies

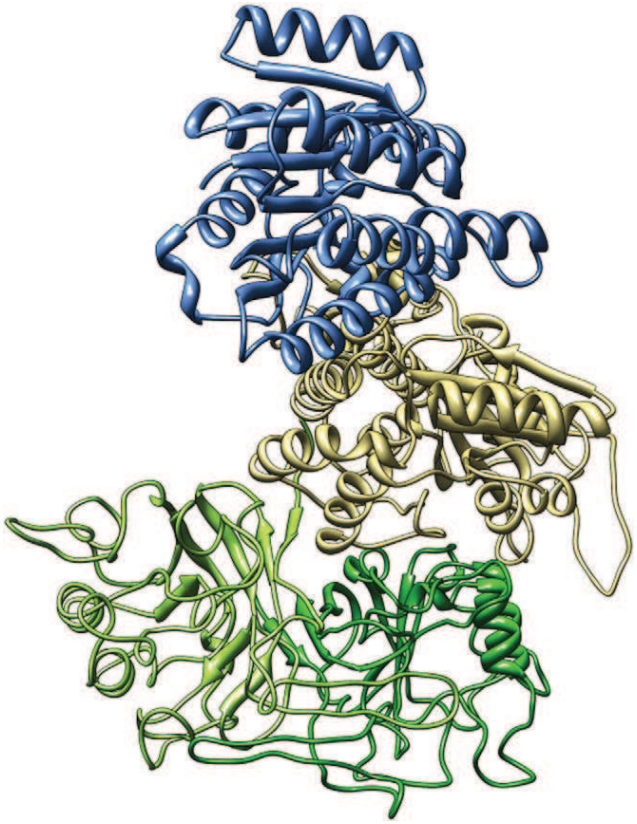

Figure 2. Superimposed Computational Docking Solutions for the Protein-Protein Interactions between DpsD/DpsE and DpsY/DpsD
DpsE (blue) docks in a region between the small subunit and helical flap of DpsD (yellow). The DpsY subunits (in different shades of green) interact planar to DpsD.

with Phe55 from the loop connecting helices 1 and 2 on DpsA. It is possible that flexible docking would show closer interactions with the conformational change in the loop connecting helices 1 and 2 in the DpsA. In the Y2H assays, KSα (DpsA) was found to establish homotypic interactions. In order to investigate whether this association would have any influence on the orientation of the interaction with MCAT (DpsD), a docking simulation was also performed between KSα (DpsA), as a homodimer, and MCAT (DpsD), and the DpsA monomer with DpsD. In both simulations, matching complexes were the first solutions from both programs.

The proposed orientation for the docking between MCAT (DpsD) and the homodimer of the CYC (DpsY) revealed an interesting interface between these two proteins (Figure 2). This was the eighth solution in Firedock and the third solution in ClusPro. The MCAT (DpsD) appeared to interact mainly via interactions established with the C terminus of helix α10 (Pro203–Ser218), the N terminus of helix α11 (Leu240–Leu252), and the loop connecting these helices (Thr219–Gln239). DpsY interacts with both of its subunits, with DpsD helix α10, penetrating into a pocket between the monomer subunits. The β1 sheet (Met6–Glu8) and loop connecting β sheets 2 and 3 from one DpsY monomer, and the loops connecting sheet β4 and helix α2, and sheet β6 and helix α3 from the second DpsY monomer are involved in the interaction.

The simulation for the docking between the MCAT and KR (DpsD/DpsE) revealed a particular interface chosen from the first solution of FireDock and the second of ClusPro with a Cα RMSD of 1.92 Å (Figure 2). MCAT (DpsD) takes part in the interaction with the helices α3 (Arg54–Asp60), α4 (Ala63–Asp67) and the loop connecting helices α2 and α3 (Asp40–Leu53). This corresponds to the region between the small subunit and helical flap of MCAT. The KR (DpsE) contributes to the interface with the C-terminal region of helix α8 (Glu205–Lys215), α6 (Ala150–Leu170), small helix α5 (Thr140–Gly142), and the loop connecting helices α5 and α6 (Lys143–Gly149). Hydrophobic amino acid residues are also likely to be involved in establishing the interaction. The simulation for the docking between the KS-CLF/CYC (DpsA-B/DpsY) showed an interesting complex where DpsY dimmer positions itself on DpsA in the proximity of the entrance to the amphipathic tunnel. The selected complex was the first solution in FireDock and the fifth in ClusPro. The DpsY dimer in this solution interacts with both KS and CLF. DpsY takes part in the interaction mostly with loops used in subunit binding that interact with helix α16 (Lys314–Tyr333), C terminal of helix α11 (Pro202–Ala210) and a loop connecting sheet β6 and helix α14 (Asn271–Gly282) on DpsA, and with helix α7 (Pro126–His128), helix α4 (Lys66–Gln71) and loop connecting helix α5 and sheet β4 (Ser90–Glu99) on DpsB. This arrangement of proteins would be favorable to reducing spontaneous cyclization, and is also in close proximity to the amphipathic tunnel (Figure 3). This represents what is, to our knowledge, the first study attempting to analyze in vivo protein interactions forming a type II PKS. A better understanding of the protein-protein interactions within the type II PKS complex should allow us to formulate new design rules for the synthesis of aromatic polyketides through combinatorial biosynthesis.

SIGNIFICANCE

Using a yeast two-hybrid (Y2H) screen, the core components forming the polyketide synthase (PKS) complex were the ketosynthase (KS) subunits, predicted to be a heterotetramer with the two KSα (DpsA) polypeptides interacting strongly with each other, and with KSβ (DpsB). The heterodimeric core was further extended to include two malonyl-CoA:ACP acyltransferase (MCAT) (DpsD) polypeptides, again interacting strongly with KSα (DpsA). Correlating our data with those of previous in vivo and in vitro experiments (Rajgarhia et al., 2001), we propose that, within the complex, the MCAT (DpsD) might act in a structural role; perhaps its physical position prevents chain initiation using an acetate starter. The cyclase (CYC) (DpsY) was found to interact with all of the proteins forming the complex, which may indicate a significant structural role, maintaining the complex in a biologically active configuration, as has been suggested for post-PKS modifying activities of other type II complexes (Petkovic et al., 1999; Perić-Concha et al., 2005). From the Y2H assays, KSα (DpsA) was predicted to play a key role in the proposed head-to-tail arrangement of the "minimal" PKS and, therefore, was chosen as the target protein to fuse to the tandem affinity purification tag. The "pulldown" experiments resulted in the purification of the KSα (DpsA) and KSβ (DpsB)

Chemistry & Biology
Type II Polyketide Synthase Structural Studies

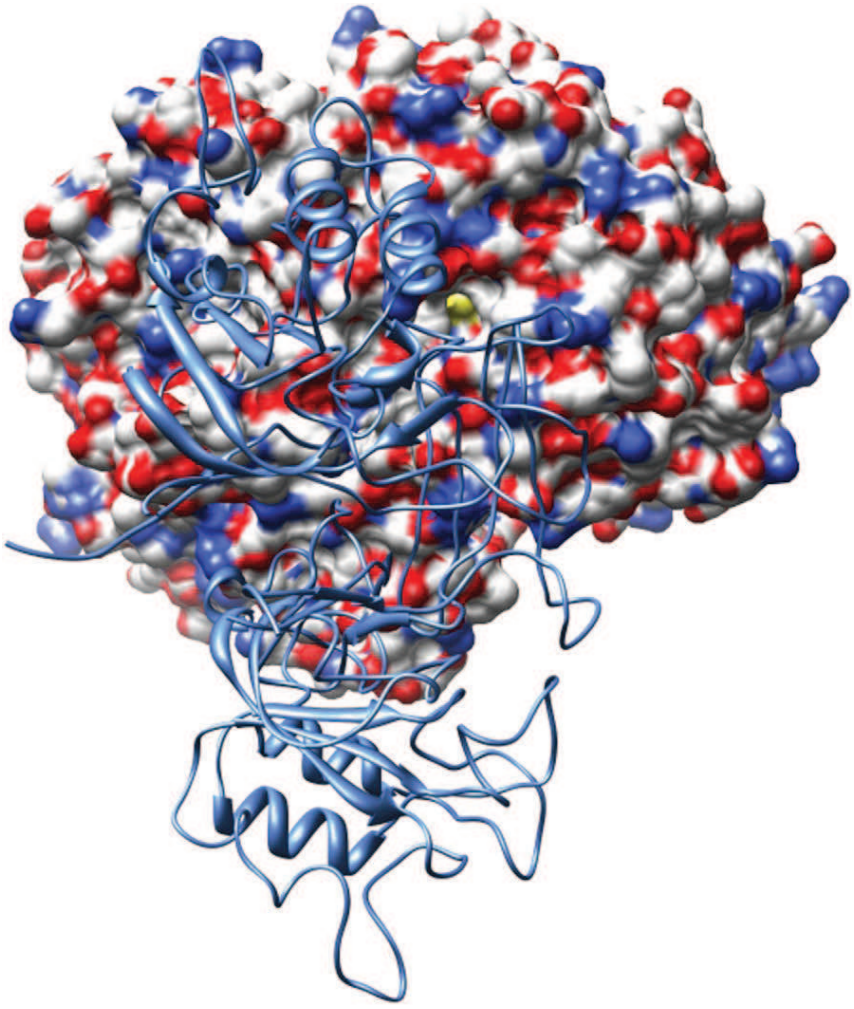

Figure 3. Computational Docking Simulating the Protein-Protein Interactions between DpsAB and DpsY

The location of the docking site is in the proximity of the entrance to the amphipathic tunnel (shown yellow on DpsAB surface). The entrance to the cavity containing active site of DpsY is also in the proximity of the entrance to the amphiphatic tunnel (Thompson et al., 2004). The surface of the DpsAB complex is colored based on hydrophobic properties of the residues using the Kyte-Doolittle scale. Colors range from blue, for the most hydrophilic residues, through white, and then to red for the most hydrophobic residues.

subunits, but failed to copurify the MCAT (DpsD), which, from the previous Y2H results, was predicted to interact strongly with KSα (DpsA). Both programs used for docking simulations were also able to predict a docking orientation for KSα/KSβ (DpsA/DpsB), similar to that observed for the solved crystal structure of the actinorhodin KSα/KSβ heterodimer (Keatinge-Clay et al., 2004). The docking simulations also suggest that the MCAT (DpsD) might have a structural role in bringing the active sites of the ketoreductase (DpsE) and the CYC (DpsY) in to close proximity, allowing the proteins to carry out adjacent modifications of the aklanonic acid backbone in conjunction with the CYC/aromatase (DpsF).

EXPERIMENTAL PROCEDURES

Y2H Experiments

Each *dps* gene was subcloned by PCR using the plasmid template pWHM1012. All the forward primers were designed incorporating an *Nde*I site upstream of the translational start codon, while all the reverse primers were designed with an EcoRI site downstream of the translational stop codon (Table S1). The amplified products were purified from agarose gels using QIAEX II following the manufacturer's instructions (QIAGEN GmbH, Hilden, Germany). The purified fragments were then cloned via NdeI/EcoRI recognition sites into the polylinker of both Y2H plasmids pGADT-7 (prey) and pGBKT-7 (bait). The constructs were sequenced to check that the *dps* gene inserts were in frame with Gal4-AD (pGADT-7) and Gal4-BD (pGBKT-7). Y2H experiments were performed in *S. cerevisiae* AH109 using procedures described by the manufacturer (BD Biosciences/Clonetech, Palo Alto, CA). Triplicate yeast transformations, using each *dps* gene in combination as both prey and bait, were plated out on both SD-2 (Trp- and Leu-) and SD-4 (Trp-, Leu-, Ade-, and His-) media and incubated at 30°C for up to 14 days. Colony lift *LacZ* assays were also performed on yeast cells grown on SD-2 medium to verify protein interactions. Protein expression was checked where no interactions were observed. Protein extraction was performed using 1 ml of SD-2 cultures (1×10^7 cells ml^{-1}) with trichloroacetic acid. Precipitated proteins were resuspended in 300 μl of SU buffer (5% w/v SDS, 8 M urea, 125 mM Tris-HCl, pH 6.8, 0.1% EDTA, 15 mg/ml DTT, and 0.005% w/v bromophenol blue). The proteins were separated by 10% SDS-PAGE and electroblotted onto nitrocellulose membranes for Western blot analysis. The antibodies used to perform Western blots were obtained from Abcam Limited (Cambridge, UK). The primary antibodies were mouse monoclonal anti-c-Myc and mouse monoclonal anti-HA tag. The secondary antibody was a rabbit polyclonal to the mouse IgG H&L horseradish peroxidase-conjugated anti-IgG. Detection was performed using 3,3′,5,5′-tetramethylbenzidine, as specified by the manufacturer (Sigma Aldrich, St. Louis, MO).

TAP

Plasmid designated pNC147 is a derivative of pWHM1012 encoding the genes *dpsA, dpsB, dpsC, dpsD*, nd *dpsG*, which translates a TAP tag fused to the N terminus of DpsA. Details of the cloning strategy to construct pNC147, in five steps, are described in Figure S2. Expression of proteins encoded by pNC147 used the heterologous host *S. coelicolor* A3(2) grown in 50 ml of YEME medium supplemented with thiostrepton (15 μg/ml) as described by Kieser et al. (2000). The mycelium was collected by centrifugation and washed twice with 0.09 M Tris-Cl, pH 7.9. The TAP tag procedure followed was as described by Rigaut et al. (1999). Proteins were separated by 1D SDS-PAGE and digested in-gel with trypsin. The peptide digests were extracted from each gel band and separated by nanoRP-LC (Micromass CapLC, Waters) before MS analysis using a Q-TOF Ultima Global (Waters). The mass spectrometric acquisition was performed in a data-dependent manner, with a 1 s MS survey scan followed by MS/MS scans (1 s) on the 3 most abundant multiply charged ions. The raw MS/MS spectra were processed to 'pkl' files using MassLynx software version 4.1 (Waters) and analyzed using GeneBio Phenyx Software (Geneva, Switzerland).

Chemistry & Biology
Type II Polyketide Synthase Structural Studies

Computer Docking Simulations

Two sets of models were built. One using Swiss-Model (Arnold et al., 2006; Guex and Peitsch, 1997) and the other using MODELER 9.4 (Fiser and Sali, 2003). Both sets were evaluated using DOPE function (Shen and Sali, 2006) from MODELER 9.4 and Verify3D (Mashiach et al., 2008). Models with the most favorable energy profile and profile score were selected for protein-protein docking simulations. For Swiss-Model 3D templates were chosen based on the results obtained using GenTHREADER (McGuffin et al., 2000), Modbase (McGuffin et al., 2000) and Predict protein (Rost et al., 2003). Templates used in MODELER were selected by scanning the query sequence against a library of sequences extracted from known structures in Protein Data Bank (PDB), which was obtained from MODELER (http://salilab.org/modeller/). Alignments were created with MODELER, unless stated otherwise. The DpsA model was built using the "Automated Mode" in Swiss-Model with 1TQYA as a template which has 62.5% sequence identity to DpsA. The DpsB model was built using the "Automated Mode" in Swiss-Model with 1TQYB as a template which has 57.75% sequence identity to DpsB. The DpsC model was built using the "Project Mode" in Swiss-Model with 1HNJA as the template which had 22.3% sequence identity to DpsC. The model was built based on an mGenTHREADER (McGuffin et al., 2000) sequence alignment and optimized using the SWISS-MODEL server (Arnold et al., 2006). The DpsD model was built using MODELER 9.4. As templates, 1MLAA (30% identity), 1NM2A (34% identity) and 2CUYA (32% identity) were used. The DpsE model was built using MODELER 9.4; 1X7GA (59% identity) and 2PH3A (41% identity) were used as templates. The DpsF model was built using MODELER 9.4 with 2RERA as the template, which had 23.2% sequence identity to DpsF. The model was built based on an mGenTHREADER sequence alignment. The DpsG model was built using MODELER 9.4; 1NQ4A (40% identity), 1OR5A (38% identity), and 1AF8A (38% identity) were used as templates. The DpsY model was built using MODELER 9.4 with 1R61A as the template, which had 24.9% sequence identity to DpsY. The model was built based on an mGenTHREADER sequence alignment; the model was built as a dimer. Sequence alignments are available upon request to the corresponding author.

These proposed models cover the KSα (DpsA) amino acid sequence from Arg3 to Arg419, with the expected structural characteristics including the position of the catalytic Cys169. The model for KSβ covers amino acid sequence from Arg26 to Ala424, with the highly conserved Gln161 of the actinorhodin homolog occupying position 183 in the primary sequence of KSβ (DpsB). The homodimer of KSα (DpsA) was obtained by docking simulation using PatchDock/FireDock. The heterodimer of DpsA/DpsB was obtained by docking simulation using PatchDock/FireDock (first solution). The putative models for the MCAT (DpsD) and KR (DpsE) cover their entire primary structure. A 3D structure for the CYC (DpsY) was modeled using the deposited structure of a predicted metal-dependent hydrolase from *Bacillus stearothermophilus* as a template (PDB accession code: 1R61). This template was chosen based on mGENthreader results. The model for the CYC (DpsY) covers its amino acid sequence from Thr12 to Glu272. Protein docking simulations between each pair of predicted 3D structures were performed using PatchDock and ClusPro. PatchDock was used with default settings and the best 100 solutions were refined using FireDock. The first 10 solutions were returned as results. ClusPro was used with default settings, and ZDOCK was used as the docking program. The top 10 solutions were returned as results. Solutions that were found by both programs were chosen for analysis of the interface. Docking simulation between the two KS subunits (DpsA and DpsB) were performed using KSα (DpsA) as receptor and KSβ (DpsB) as ligand, based upon the published crystal structure for the actinorhodin orthologs. Docking simulation between KSα/CYC (DpsA/DpsY) were performed using the dimer of the CYC (DpsY) as receptor and KSα (DpsA) as ligand. A docking simulation between KSβ and CYC (DpsB/DpsY) was performed in a similar fashion. The docking between the two subunits of the KS dimer (DpsA/DpsB) and the homodimer of the CYC (DpsY) was performed using the KSα/KSβ (DpsA/DpsB) dimer as receptor and the CYC (DpsY) homodimer as ligand. Docking simulation between the KSα and MCAT (DpsA/DpsD) was performed using KSα (DpsA) as receptor and MCAT (DpsD) as ligand. Docking between homodimer of KSα (DpsA) and MCAT (DpsD) was also performed using the KSα (DpsA) homodimer as receptor and MCAT (DpsD) as ligand. For the KR/CYC (DpsE/DpsY) simulation, the homodimer of the CYC (DpsY) was the receptor and KR (DpsE) was the ligand. The MCAT/CYC (DpsD/DpsY) docking was simulated using the homodimer of the CYC (DpsY) as receptor and MCAT (DpsD) as ligand. The MCAT/KR (DpsD/DpsE) docking was simulated using MCAT (DpsD) as receptor and KR (DpsE) as ligand.

SUPPLEMENTAL DATA

Supplemental Data include one table and seven figures and can be found with this article online at http://www.chembiol.com/cgi/content/full/15/11/1156/DC1/.

ACKNOWLEDGMENTS

This work was supported by Wellcome Trust grant 064197 to P.F.L., the Central Research Fund, University of London (ref: AR/CRF/B), by Ministry of Science, Education, and Sports, Republic of Croatia grant TP-05/0058-23 to D.H. and J.Z., and by a stipend from the School of Pharmacy, University of London, to G.C. and S.H. The authors would like to thank M. Smith and F.A. Stephenson for their technical support and advice.

Received: March 28, 2008
Revised: August 9, 2008
Accepted: September 4, 2008
Published: November 21, 2008

REFERENCES

Andrusier, N., Nussinov, R., and Wolfson, H.J. (2007). FireDock: fast interaction refinement in molecular docking. Proteins *69*, 139–159.

Arnold, K., Bordoli, L., Kopp, J., and Schwede, T. (2006). The SWISS-MODEL Workspace: A web-based environment for protein structure homology modelling. Bioinformatics *22*, 195–201.

Bairoch, A., Apweiler, R., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., et al. (2005). The Universal Protein Resource (UniProt). Nucleic Acids Res. *33*, D154–D159.

Bao, W., Sheldon, P.J., and Hutchinson, C.R. (1999a). Purification and properties of the *Streptomyces peucetius* DpsC beta-ketoacyl:acyl carrier protein synthase III that specifies the propionate-starter unit for type II polyketide biosynthesis. Biochemistry *38*, 9752–9757.

Bao, W., Sheldon, P.J., Wendt-Pienkowski, E., and Hutchinson, C.R. (1999b). The *Streptomyces peucetius* dpsC gene determines the choice of starter unit in biosynthesis of the daunorubicin polyketide. J. Bacteriol. *181*, 4690–4695.

Beinker, P., Lohkamp, B., Peltonen, T., Niemi, J., Mantsala, P., and Schneider, G. (2006). Crystal structures of SnoaL2 and AclR: two putative hydroxylases in the biosynthesis of aromatic polyketide antibiotics. J. Mol. Biol. *359*, 728–740.

Bibb, M.J., Sherman, D.H., Omura, S., and Hopwood, D.A. (1994). Cloning, sequencing and deduced functions of a cluster of *Streptomyces* genes probably encoding biosynthesis of the polyketide antibiotic frenolicin. Gene *142*, 31–39.

Bisang, C., Long, P.F., Cortés, J., Westcott, J., Crosby, J., Matharu, A.-L., Cox, R.J., Simpson, T.J., Staunton, J., and Leadlay, P.F. (1999). A chain initiation factor common to both aromatic and modular polyketide synthases. Nature *401*, 502–505.

Carreras, C.W., and Khosla, C. (1998). Purification and *in vitro* reconstitution of the essential protein components of an aromatic polyketide synthase. Biochemistry *37*, 2084–2088.

Castaldo, G., Crosby, J., and Long, P.F. (2005). Exploring protein interactions on a minimal type II polyketide synthase using a yeast two-hybrid system. Food Technol. Biotechnol. *43*, 109–112.

Chen, R., Li, L., and Weng, Z. (2003). ZDOCK: an initial-stage protein-docking algorithm. Proteins *52*, 80–87.

Comeau, S.R., Gatchell, D.W., Vajda, S., and Camacho, C.J. (2003). ClusPro: an automated docking and discrimination method for the prediction of protein complexes. Bioinformatics *20*, 45–50.

Crump, M.P., Crosby, J., Dempsey, C.E., Parkinson, J.A., Murray, M., Hopwood, D.A., and Simpson, T.J. (1997). Solution structure of the actinorhodin

polyketide synthase acyl carrier protein from *Streptomyces coelicolor* A3(2). Biochemistry *36*, 6000–6008.

Decker, H., Summers, R.G., and Hutchinson, C.R. (1994). Overproduction of the acyl carrier protein component of a type II polyketide synthase stimulates production of tetracenomycin biosynthetic intermediates in *Streptomyces glaucescens*. J. Antibiot. (Tokyo) *47*, 54–63.

Dreier, J., and Khosla, C. (2000). Mechanistic analysis of a type II polyketide synthase. Role of conserved residues in the beta-ketoacyl synthase-chain length factor heterodimer. Biochemistry *39*, 2088–2095.

Duhovny, D., Nussinov, R., and Wolfson, H.J. (2002). Efficient unbound docking of rigid molecules. In Proceedings of the 2nd Workshop on Algorithms in Bioinformatics (WABI), R. Guigo and D. Gusfield, eds. (New York: Springer), 185–200.

Findlow, S.C., Winsor, C., Simpson, T.J., Crosby, J., and Crump, M.P. (2003). Solution structure and dynamics of oxytetracycline polyketide synthase acyl carrier protein from *Streptomyces rimosus*. Biochemistry *42*, 8423–8433.

Fiser, A., and Sali, A. (2003). Comparative protein structure modeling with MODELLER: a practical approach. Methods Enzymol. *374*, 463–493.

Florova, G., Kazanina, G., and Reynolds, K.A. (2002). Enzymes involved in fatty acid and polyketide biosynthesis in *Streptomyces glaucescens*: role of FabH and FabD and their acyl carrier protein specificity. Biochemistry *41*, 10462–10471.

Funa, N., Ohnishi, Y., Fujii, I., Shibuya, M., Ebizuka, Y., and Horinouchi, S. (1999). A new pathway for polyketide synthesis in microorganisms. Nature *400*, 897–899.

Gramajo, H.C., White, J., Hutchinson, C.R., and Bibb, M.J. (1991). Overproduction and localization of components of the polyketide synthase of *Streptomyces glaucescens* involved in the production of the antibiotic tetracenomycin C. J. Bacteriol. *173*, 6475–6483.

Grimm, A., Madduri, K., Ali, A., and Hutchinson, C.R. (1994). Characterization of the *Streptomyces peucetius* ATCC 29050 genes encoding doxorubicin polyketide synthase. Gene *151*, 1–10.

Guex, N., and Peitsch, M.C. (1997). SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modelling. Electrophoresis *18*, 2714–2723.

Hadfield, A.T., Limpkin, C., Teartasin, W., Simpson, T., Crosby, J., and Crump, M.P. (2004). The crystal structure of the actIII actinorhodin polyketide reductase: proposed mechanism for ACP and polyketide binding. Structure *12*, 1865–1875.

Hertweck, C., Luzhetskyy, A., Rebets, Y., and Bechtold, A. (2007). Type II polyketide synthases: gaining a deeper insight into enzymatic teamwork. Nat. Prod. Rep. *24*, 162–190.

Hesketh, A.R., Chandra, G., Shaw, A.D., Rowland, J.J., Kell, D.B., Bibb, M.J., and Chater, K.F. (2002). Primary and secondary metabolism, and post-translational protein modifications, as portrayed by proteomic analysis of *Streptomyces coelicolor*. Mol. Microbiol. *46*, 917–932.

Hitchman, T.S., Crosby, J., Byrom, K.J., Cox, R.J., and Simpson, T.J. (1998). Catalytic self-acylation of type II polyketide synthase acyl carrier proteins. Chem. Biol. *5*, 35–47.

Hutchinson, C.R., and Colombo, A.L. (1999). Genetic engineering of doxorubicin production in *Streptomyces peucetius*: a review. J. Ind. Microbiol. Biotechnol. *23*, 647–652.

Jackowski, S., Murphy, C.M., Cronan, J.E., Jr., and Rock, C.O. (1989). Acetoacetyl-acyl carrier protein synthase. A target for the antibiotic thiolactomycin. J. Biol. Chem. *264*, 7624–7629.

Jansson, A., Koskiniemi, H., Mäntsälä, P., Niemi, J., and Schneider, G. (2004). Crystal structure of a ternary complex of DnrK, a methyltransferase in daunorubicin biosynthesis, with bound products. J. Biol. Chem. *279*, 41149–41156.

Kealey, J.T., Liu, L., Santi, D.V., Betlach, M.C., and Barr, P.J. (1998). Production of a polyketide natural product in nonpolyketide-producing prokaryotic and eukaryotic hosts. Proc. Natl. Acad. Sci. USA *95*, 505–509.

Keatinge-Clay, A.T., Shelat, A.A., Savage, D.F., Tsai, S.C., Miercke, L.J., O'Connell, J.D., Khosla, C., and Stroud, R.M. (2003). Catalysis, specificity, and ACP docking site of *Streptomyces coelicolor* malonyl-CoA:ACP transacylase. Structure *11*, 147–154.

Keatinge-Clay, A.T., Maltby, D.A., Medzihradszky, K.F., Khosla, C., and Stroud, R.M. (2004). An antibiotic factory caught in action. Nat. Struct. Mol. Biol. *11*, 888–893.

Kieser, T., Bibb, M.J., Buttner, M.J., and Hopwood, D.A. (2000). Practical *Streptomyces* Genetics (Norwich, UK: The John Innes Foundation).

Korman, T.P., Hill, J.A., Vu, T.N., and Tsai, S.C. (2004). Structural analysis of actinorhodin polyketide ketoreductase: cofactor binding and substrate specificity. Biochemistry *43*, 14529–14538.

Kramer, P.J., Zawada, R.J.X., McDaniel, R., Hutchinson, C.R., Hopwood, D.A., and Khosla, C. (1997). Rational design and engineered biosynthesis of a novel 18-carbon aromatic polyketide. J. Am. Chem. Soc. *119*, 635–639.

Li, Q., Khosla, C., Puglisi, J.D., and Liu, C.W. (2003). Solution structure and backbone dynamics of the holo form of the frenolicin acyl carrier protein. Biochemistry *42*, 4648–4657.

Lomovskaya, N., Doi-Katayama, Y., Filippini, S., Nastro, C., Fonstein, L., Gallo, M., Colombo, A.L., and Hutchinson, C.R. (1998). The *Streptomyces peucetius dpsY* and *dnrX* genes govern early and late steps of daunorubicin and doxorubicin biosynthesis. J. Bacteriol. *180*, 2379–2386.

Mashiach, E., Schneidman-Duhovny, D., Andrusier, N., Nussinov, R., and Wolfson, H.J. (2008). FireDock: a Web server for fast interaction refinement in molecular docking. Nucleic Acids Res. *36*, W229–W232.

Matharu, A.L., Cox, R.J., Crosby, J., Byrom, K.J., and Simpson, T.J. (1998). MCAT is not required for in vitro polyketide synthesis in a minimal actinorhodin polyketide synthase from *Streptomyces coelicolor*. Chem. Biol. *5*, 699–711.

McDaniel, R., Ebert-Khosla, S., Hopwood, D.A., and Khosla, C. (1995). Rational design of aromatic polyketide natural products by recombinant assembly of enzymatic subunits. Nature *375*, 549–554.

McGuffin, L.J., Bryson, K., and Jones, D. (2000). The PSIPRED protein structure prediction server. Bioinformatics *16*, 404–405.

Meurer, G., Gerlitz, M., Wendt-Pienkowski, E., Vining, L.C., Rohr, J., and Hutchinson, C.R. (1997). Iterative type II polyketide synthases, cyclases and ketoreductases exhibit context-dependent behavior in the biosynthesis of linear and angular decapolyketides. Chem. Biol. *4*, 433–443.

Minotti, G., Menna, P., Salvatorelli, E., Cairo, G., and Gianni, L. (2004). Anthracyclines: molecular advances and pharmacologic developments in antitumor activity and cardiotoxicity. Pharmacol. Rev. *56*, 185–229.

Mootz, H.D., Finking, R., and Marahiel, M.A. (2001). 4′-phosphopantetheine transfer in primary and secondary metabolism of *Bacillus subtilis*. J. Biol. Chem. *276*, 37289–37298.

Moult, J. (2005). A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. Curr. Opin. Struct. Biol. *15*, 285–289.

Pan, H., Tsai, S., Meadows, E.S., Miercke, L.J., Keatinge-Clay, A.T., O'Connell, J., Khosla, C., and Stroud, R.M. (2002). Crystal structure of the priming beta-ketosynthase from the R1128 polyketide biosynthetic pathway. Structure *10*, 1559–1568.

Parrish, J.R., Gulyas, K.D., and Finley, R.L., Jr. (2006). Yeast two-hybrid contributions to interactome mapping. Curr. Opin. Biotechnol. *17*, 387–393.

Perić-Concha, N., Borovička, B., Long, P.F., Hranueli, D., Waterman, P.G., and Hunter, I.S. (2005). Ablation of the *otc*C gene encoding a post-polyketide hydroxylase from the oxytetracyline biosynthetic pathway in *Streptomyces rimosus* results in novel polyketides with altered chain length. J. Biol. Chem. *280*, 37455–37460.

Petkovic, H., Thamchaipenet, A., Zhou, L.H., Hranueli, D., Raspor, P., Waterman, P.G., and Hunter, I.S. (1999). Disruption of the aromatase/cyclase from the oxytetracycline gene cluster of *Streptomyces rimosus* results in production of novel polyketides with shorter chain lengths. J. Biol. Chem. *274*, 32829–32834.

Piel, J., Hertweck, C., Shipley, P.R., Hunt, D.M., Newman, M.S., and Moore, B.S. (2000). Cloning, sequencing and analysis of the enterocin biosynthesis gene cluster from the marine isolate '*Streptomyces maritimus*': evidence for the derailment of an aromatic polyketide synthase. Chem. Biol. *7*, 943–955.

Chemistry & Biology
Type II Polyketide Synthase Structural Studies

Qin, J., Vinogradova, O., and Gronenborn, A.M. (2001). Protein-protein interactions probed by nuclear magnetic resonance spectroscopy. Methods Enzymol. *339*, 377–389.

Qiu, X., Janson, C.A., Smith, W.W., Head, M., Lonsdale, J., and Konstantinidis, A.K. (2001). Refined structures of β-ketoacyl-acyl carrier protein synthase III. J. Mol. Biol. *307*, 341–356.

Rajgarhia, V.B., and Strohl, W.R. (1997). Minimal *Streptomyces* sp. strain C5 daunorubicin polyketide biosynthesis genes required for aklanonic acid biosynthesis. J. Bacteriol. *179*, 2690–2696.

Rajgarhia, V.B., Priestley, N.D., and Strohl, W.R. (2001). The product of dpsC confers starter unit fidelity upon the daunorubicin polyketide synthase of *Streptomyces* sp. strain C5. Metab. Eng. *3*, 49–63.

Raty, K., Kantola, J., Hautala, A., Hakala, J., Ylihonko, K., and Mantsala, P. (2002). Cloning and characterization of *Streptomyces galilaeus* aclacinomycins polyketide synthase (PKS) cluster. Gene *293*, 115–122.

Revill, W.P., Bibb, M.J., and Hopwood, D.A. (1995). Purification of a malonyltransferase from *Streptomyces coelicolor* A3(2) and analysis of its genetic determinant. J. Bacteriol. *177*, 3946–3952.

Rigaut, G., Shevchenko, A., Rutz, B., Wilm, M., Mann, M., and Séraphin, B. (1999). A generic protein purification method for protein complex characterization and proteome exploration. Nat. Biotechnol. *17*, 1030–1032.

Rost, B., Yachdav, G., and Liu, J. (2003). The PredictProtein server. Nucleic Acids Res. *32*, W321–W326. 10.1093/nar/gkh377.

Schneidman-Duhovny, D., Inbar, Y., Nussinov, R., and Wolfson, H.J. (2005). PatchDock and SymmDock: servers for rigid and symmetric docking. Nucleic Acids Res. *33*, W363–W367. 10.1093/nar/gki481.

Shen, M.Y., and Sali, A. (2006). Statistical potential for assessment and prediction of protein structures. Protein Sci. *15*, 2507–2524.

Sultana, A., Kallio, P., Jansson, A., Wang, J.S., Niemi, J., Mäntsälä, P., and Schneider, G. (2004). Structure of the polyketide cyclase *SnoaL* reveals a novel mechanism for enzymatic aldol condensation. EMBO J. *23*, 1911–1921.

Summers, R.G., Ali, A., Shen, B., Wessel, W.A., and Hutchinson, C.R. (1995). Malonyl-coenzyme A: acyl carrier protein acyltransferase of *Streptomyces glaucescens*: A possible link between fatty acid and polyketide biosynthesis. Biochemistry *34*, 9389–9402.

Tang, Y., Lee, T.S., Kobayashi, S., and Khosla, C. (2003). Ketosynthases in the initiation and elongation modules of aromatic polyketide synthases have orthogonal acyl carrier protein specificity. Biochemistry *42*, 6588–6595.

Tang, Y., Lee, T.S., and Khosla, C. (2004). Engineered biosynthesis of regioselectively modified aromatic polyketides using bimodular polyketide synthases. PLoS Biol. *2*, E31. 10.1371/journal.pbio.0020031.

Thompson, T.B., Katayama, K., Watanabe, K., Hutchinson, C.R., and Rayment, I. (2004). Structural and functional analysis of tetracenomycin F 2 cyclase from *Streptomyces glaucescens*: a type II polyketide cyclase. J. Biol. Chem. *279*, 37956–37963.

Tramontano, A., and Morea, V. (2003). Assessment of homology-based predictions in CASP5. Proteins *53*, 352–368.

Wattanachaisaereekul, S., Lantz, A.E., Nielsen, M.L., and Nielsen, J. (2008). Production of the polyketide 6-MSA in yeast engineered for increased malonyl-CoA supply. Metab. Eng. Epub ahead of print May 10, 2008. 10.1016/j.ymben.2008.04.005.

White, S.W., Zheng, J., Zhang, Y.M., and Rock, C.O. (2005). The structural biology of type II fatty acid biosynthesis. Annu. Rev. Biochem. *74*, 791–831.

Wohlert, S.E., Wendt-Pienkowski, E., Bao, W., and Hutchinson, C.R. (2001). Production of aromatic minimal polyketides by the daunorubicin polyketide synthase genes reveals the incompatibility of the heterologous DpsY and JadI cyclases. J. Nat. Prod. *64*, 1077–1080.

Yu, T.-W., Shen, Y., McDaniel, R., Floss, H.G., Khosla, C., Hopwood, D.A., and Moore, B.S. (1998). Engineered biosynthesis of novel polyketides from *Streptomyces* spore pigment polyketide synthases. J. Am. Chem. Soc. *120*, 7749–7759.

Zuiderweg, E.R. (2002). Mapping protein-protein interactions in solution by NMR spectroscopy. Biochemistry *41*, 1–7.

2.4 Polyketide synthase genes and the natural products potential of Dictyostelium discoideum

Jurica Zucko, Nives Skunca, Tomaz Curk, Blaz Zupan, Paul F. Long, John Cullum, Richard Kessin and Daslav Hranueli. *Bioinformatics*, ***23***, 2543–2549, 2007.

Abstract:

Motivation: The genome of the social amoeba *Dictyostelium discoideum* contains an unusually large number of polyketide synthase (PKS) genes. An analysis of the genes is a first step towards understanding the biological roles of their products and exploiting novel products.

Results: 45 Type I iterative PKS genes were found, 5 of which are probably pseudogenes. Catalytic domains that are homologous with known PKS sequences as well as possible novel domains were identified. The genes often occurred in clusters of 2-5 genes, where members of the cluster had very similar sequences. The *D. discoideum* PKS genes formed a clade distinct from fungal and bacterial genes. All nine genes examined by RT-PCR were expressed, although at different developmental stages. The promoters of PKS genes were much more divergent than the structural genes, although we have identified motifs that are unique to some PKS gene promoters.

Own contribution to the paper:

Annotation and reconstruction of polyketide genes in *Dictyostelium discoideum* using custom-made HMM profiles. Phylogenetic analysis of PKS genes and domains and supervision of an undergraduate student (Miss. Nives Škunca)

BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 19 2007, pages 2543–2549
doi:10.1093/bioinformatics/btm381

Sequence analysis

Polyketide synthase genes and the natural products potential of *Dictyostelium discoideum*

J. Zucko[1,5], N. Skunca[1], T. Curk[2], B. Zupan[2,3], P.F. Long[4], J. Cullum[5], R.H. Kessin[6] and D. Hranueli[1,*]

[1]Faculty of Food Technology and Biotechnology, University of Zagreb, Pierottijeva 6, 10000 Zagreb, Croatia, [2]Faculty of Computer and Information Science, University of Ljubljana, Tržaška cesta 25, SI-1001 Ljubljana, Slovenia, [3]Department of Molecular and Human Genetics, Baylor College of Medicine, 1 Baylor Plaza, Houston, TX 77030, USA, [4]The School of Pharmacy, University of London, 29/39 Brunswick Square, London WC1N 1AX, UK, [5]Department of Genetics, University of Kaiserslautern, Postfach 3049, 67653 Kaiserslautern, Germany and [6]Department of Anatomy and Cell Biology, Columbia University, 630 West 168th Street, 10032 New York, USA

Received on May 15, 2007; revised on July 9, 2007; accepted on July 10, 2007

Advance Access publication July 27, 2007

Associate Editor: Dmitrij Frishman

ABSTRACT

Motivation: The genome of the social amoeba *Dictyostelium discoideum* contains an unusually large number of polyketide synthase (PKS) genes. An analysis of the genes is a first step towards understanding the biological roles of their products and exploiting novel products.

Results: A total of 45 Type I iterative PKS genes were found, 5 of which are probably pseudogenes. Catalytic domains that are homologous with known PKS sequences as well as possible novel domains were identified. The genes often occurred in clusters of 2–5 genes, where members of the cluster had very similar sequences. The *D.discoideum* PKS genes formed a clade distinct from fungal and bacterial genes. All nine genes examined by RT–PCR were expressed, although at different developmental stages. The promoters of PKS genes were much more divergent than the structural genes, although we have identified motifs that are unique to some PKS gene promoters.

Contact: dhranueli@pbf.hr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

The amoebae of *Dictyostelium discoideum* live in the soil and feed on a variety of bacteria and fungi. When the food is exhausted, the amoebae collect into mounds and then produce fruiting bodies. Many laboratories study the chemotaxis, the cell motility and the differentiation that are involved in fruiting body formation (Kessin, 2001). The genetics of the organism is well developed and it is possible to introduce genes and to knock-out genes. The genome sequence was recently completed showing about 12 500 genes in a relatively small AT-rich genome of 34 Mb (Eichinger *et al.*, 2005). The organism is exceptionally rich in polyketide synthases (PKS) with 43 putative genes spread singly and in clusters on all six chromosomes being reported in the initial annotation (Eichinger *et al.*, 2005). The only known polyketide product is differentiation inducing factor (DIF), which induces a particular subset of stalk cells during the complex development of the organism. The laboratory of Rob Kay characterized DIF and showed that its PKS is unusual in possessing a novel chalcone synthase domain (Austin *et al.*, 2006 and references therein). DIF has also been suggested to have mitochondrial uncoupling (Shaulsky and Loomis, 1995) and antiproliferation properties (Akaishi *et al.*, 2004; Kubohara *et al.*, 2003). There may well be further PKS genes involved in cell–cell communication, but it is also likely that some of them encode products to achieve competitive advantages in the soil (e.g. antibiotics). Polyketides are ubiquitous in nature and have been isolated from microorganisms, plants and invertebrates. They have found widespread use in the pharmaceutical, agrochemical and biotechnology industries. PKSs are large multienzyme protein complexes that contain a coordinated group of catalytic sites (Hranueli *et al.*, 2005). Type I PKSs are multifunctional proteins composed of all the active sites required for polyketide biosynthesis, which are contained in a series of domains in the proteins. Biosynthesis occurs as a stepwise process using simple carboxylic acid CoA esters as substrates. The minimal requirements for a PKS are the three domains: acyl carrier protein (ACP), acyltransferase (AT) and ketosynthase (KS). These are usually present in the order KS-AT-ACP in the protein and result in incorporation of a keto group. However, there are often one or more additional reduction domains present between AT and ACP: ketoreductase (KR), dehydratase (DH) and enoyl reductase (ER), successively reduce the keto group to a hydroxyl group, a C=C double bond or a fully reduced product. Fatty acid synthases (FAS) belong to the PKS family and have all three reduction activities (Schweizer and Hofmann, 2004). When present, the reductive domains are in the order DH-ER-KR. There may also be a methyl transferase

*To whom correspondence should be addressed.

domain (MT) between DH and ER. It is common for inactive domains to be present, so that the presence of a domain does not prove that the corresponding activity occurs. The possibility to undergo none, one or more of these reduction reactions at each biosynthesis step contributes to the huge structural diversity seen in this class of natural products. The level of chemical complexity is further increased by incorporating stereo-isomers of different starter and extender units, as well as post PKS modifications, such as glycosylation, hydroxylation or methylation (Weissman and Leadlay, 2005). In bacterial systems, modular PKSs are common in which each biosynthetic step is carried out by a different module with its own KS-AT-ACP domains, which results in very large multi-modular proteins, e.g. for erythromycin biosynthesis (Khosla *et al*., 2007). In contrast, most fungal Type I systems are iterative so that a single module carries out several biosynthesis steps, e.g. for lovostatin biosynthesis (Schumann and Hertweck, 2006). In this article, we extend the analysis performed for this gene family by Eichinger *et al*. (2005). In collaboration with the curators at dictyBase (Chisholm *et al*., 2006), we have characterized the PKS genes of *D.discoideum* and shown that they are probably Type I iterative PKSs. However, many of the genes contain extra sequences that may be novel domains and there is an intriguing clustering of closely related genes in the chromosome. The PKS genes are differentially regulated, which is mirrored by divergence of the promoter regions.

2 METHODS

2.1 Identification and annotation of PKS genes

dictyBase version 2.5 (http://dictybase.org/; Chisholm *et al*., 2006) was used as a starting point for the analysis. The DNA sequences of the six chromosomes were translated using Transeq (Rice *et al*., 2000). BLAST searches used the NCBI server (http://www.ncbi.nlm.nih.gov/BLAST/; Altschul *et al*., 1990). Multiple alignments used the Clustal W service at EBI (http://www.ebi.ac.uk/clustalw; Thompson *et al*., 1994). Bacterial PKS domain sequences were obtained from the NRPS-PKS database (http://203.90.127.50/nrps-pks.html; Ansari *et al*., 2004). For profile analysis, HMMER version 2.3.2 (http://hmmer.janelia.org/; Eddy, 1998) and release 20 of the Pfam database (http://www.sanger.ac.uk/Software/Pfam/; Bateman *et al*., 2002) were used. Artemis (Rutherford *et al*., 2000) was used for annotation.

2.2 Phylogenetic analysis

KS and AT domain protein sequences were obtained for bacterial (Jenke-Kodama *et al*., 2005) and fungal PKSs (Ansari *et al*., 2004). Phylogenetic analysis was performed using the MEGA3 software package (Kumar *et al*., 2004). Uniform substitution rates at all sites were assumed and 100 or 500 replicates were used for bootstrapping. Trees were constructed with distance methods (neighbour-joining or minimal evolution) with two choices of distance model (Poisson correction or the Jones–Taylor–Thornton model). In addition, maximum parsimony was also used. Bacterial *fabF* and *fabD* genes or the chicken and human FAS genes were used as outgroups to root the trees. The sequences used are in the Supplementary Material.

2.3 Analysis of promoter regions

Motifs were detected with the program MEME (Bailey and Elkan, 1994). Motif combinations specific for PKS genes were detected using the program MAST (Bailey and Elkan, 1994). The selected motifs were matched with known motifs in the TRANSFAC (Wingender *et al*., 1996) database and sequence logos (Schneider and Stephens, 1990) were constructed. The program AlignACE (Hughes *et al*., 2000; Roth *et al*., 1998) was also used to find motifs.

2.4 PKS gene expression

Cell harvesting and RNA extraction were carried out with TRIZOL according to the recommendations of the manufacturer and as described in Van Driessche *et al*. (2005). Quantitative RT–PCR experiments were performed using syber-green real-time PCR on an Opticon system as described in Huang *et al*. (2006). Results were normalized to the transcripts of a constitutive gene, *IG7*. Gene-specific primer pairs were constructed for the following genes: *stlA* (*pks1*), *pks2*, *pks3*, *pks10*, *pks18*, *pks24*, *pks25*, *pks26*, *stlB* (*pks37*). The primer sequences are given in Supplementary Material.

3 RESULTS

3.1 Identification and domain structure of PKS genes

In dictyBase version 2.5, there were 46 sequences annotated as putative PKS genes as well as two genes annotated as putative fatty acid synthases. For our studies, the DNA sequence of each chromosome was translated in all six reading frames. Many of the putative PKS genes could be identified using BLAST with a standard KS domain (module 4 of erythromycin), but this did not show if a complete PKS gene was present. Therefore, HMM-profiles were constructed for the domains KS, AT, DH, ER, KR and ACP starting from well-characterized domains from bacterial modular PKS clusters and the sequences were analysed using the HMMER program package. The occurrence of typical PKS domains in the expected order adjacent to each other indicated the presence of a PKS gene. The newly identified domains from *D.discoideum* were used to refine the HMM-profiles to improve identification of PKS genes. This was particularly important for the DH domains, where the refined profiles recognized domains of the expected size in most of the genes, whereas the initial profile missed many DH domains. Methyl transferase (MT) domains are common in fungal PKSs (Schumann and Hertweck, 2006). MT domains were identified in many of the genes using an HMM-profile from the Pfam database (accession number PF08242). This initial analysis identified 45 putative PKS genes (*pks1–pks45*) (Supplementary Material). We did not rename the previously designated genes *stlA* (corresponding to *pks1*) or *stlB* (corresponding to *pks37*). This analysis showed considerable differences to the annotation in dictyBase version 2.5. In 14 cases, the previous annotation had recognized short sequences with resemblance to single PKS domains, but they did not have the structure of complete genes. In 12 cases, two different parts of a single PKS gene had been annotated as distinct PKS genes. There were 13 genes that had not been previously recognized as PKS genes and four new PKS genes. Although this analysis identified putative PKS genes, some sequences contained stop codons or had split domains into coding regions in different reading frames. This could either be because a pseudogene was present or because an intron had not been detected during the initial annotation. The predicted protein sequences were aligned with known PKS protein sequences using the CLUSTAL W program. In some

Fig. 1. Deduced protein structure of a typical PKS gene (*pks15*). The gene was not identified in the original genome annotation reported in dictyBase version 2.5. It is 9.453 kb in size and located on chromosome 2. In addition to the known domains, there are three amino acid regions of unknown function between the DH and MT domains, between the MT and ER domains and after ACP.

cases, it was clear that plausible intron splice sites would give rise to a protein sequence in good alignment with known domains. Forty of the 45 PKS genes had a structure compatible for expression. The five genes *pks4*, *pks11*, *pks12*, *pks20* and *pks43* are probably pseudogenes as they contain stop codons that cannot easily be explained by the presence of introns and in two cases also have an aberrant structure: *pks4* lacks an ACP domain and has incomplete ER and KR domains, while *pks12* has a 500 bp inversion with an adjacent 200 bp deletion. These changes have been coordinated with the curators of dictyBase, who have access to the original sequencing reads, and are now included in this online resource (Chisholm *et al*., 2006).

Among the 40 probably expressed genes, 2 were the previously described *stlA* (*pks1*) and *stlB* (*pks37*), which have chalcone-like domains. The other 38 genes all have the following set of predicted domains: KS, AT, ER, KR and ACP. Thirty seven of these genes have a DH domain of the correct length (Supplementary Material), whereas *pks21* has a short DH domain, which is probably not functional (94 instead of 154–187 amino acids in the other cases). Most genes (30/38) also contained an MT domain. Two genes that lacked the MT domain were *pks16* and *pks17*, which were suggested to be FAS. The DH and MT domains lie between AT and ER. However, in all 38 PKS genes there were also one or two substantial regions (286–838 amino acids) of no known function between AT and ER. This is shown for a typical gene (*pks15*) in Figure 1. Some of the sequences in this region are conserved between PKS genes. However, they do not give significant hits with BLAST to sequences in other organisms. Some of the sequences contribute to profiles in the Pfam-B database, but all members of the families are in the *D.discoideum* genome. Most known PKS proteins in other organisms end with the ACP domain. It was striking that 17 of the 38 predicted PKS proteins in *D.discoideum* had 288–501 additional amino acids after the ACP domain (Fig. 1 and Supplementary Material). These additional C-terminal regions did not resemble the chalcone-like domains at the C-terminals of *stlA* and *stlB*. Although there are conserved sequences between some of the genes, there was no detectable similarity to proteins in other organisms. The detailed structures of each PKS gene are given in the Supplementary Material. No Type II or modular Type I PKS genes were found. No non-ribosomally encoded peptide (NRPS) genes were found using BLAST or appropriate profiles.

3.2 Phylogeny of *D.discoideum* KS and AT domains and gene clusters

The amino acid sequences of the KS domains of the *D.discoideum* PKS genes were aligned with selected bacterial and fungal domains and used to construct a phylogenetic tree (Fig. 2) with neighbour-joining method using the Poisson correction distance model. Different tree construction methods were tested (Jones–Taylor–Thornton distance model, minimal evolution, maximum parsimony), but they did not result in significant changes in the tree (see Supplementary Material). The 45 *D.discoideum* PKS genes formed a clade (bootstrap value of 92%) distinct from the bacterial and fungal sequences. The genes *stlA* (*pks1*), *stlB* (*pks37*) and the two putative fatty acid synthase genes (*pks16* and *pks17*) were distant from the other 41 genes, which formed a clade (bootstrap value 91%). A phylogenetic tree was also constructed for the AT domains of the *D.discoideum* PKS genes (Supplementary Material). This tree showed an almost identical branching of the PKS genes compared to the KS tree. The *D.discoideum* AT-domains grouped with domains that incorporate C2 building blocks. Examination of the sequences showed that they contained C2-specific motifs (Haydock *et al*., 1995).

The protein sequences of the 38 *D.discoideum* PKS genes that are probably functional and possess all the domains KS, AT, DH, ER, KR and ACP were aligned with selected bacterial and fungal sequences that posses all the domains and used to construct a phylogenetic tree. The use of different tree construction methods (as for the KS trees) did not result in significant changes in the tree (see Supplementary Material). The *D.discoideum* PKS genes formed a clade distinct from the bacterial and fungal sequences. The phylogeny of the whole genes was little different from that of the KS and AT domains alone.

The genes are distributed over the six chromosomes. However, many of the genes are clustered. There are 10 pairs of genes (*pks11*/*pks12*, *pks16*/*pks17*, *pks22*/*pks23*, *pks24*/*pks25*, *pks27*/*pks28*, *pks33*/*pks34*, *pks35*/*pks36*, *pks38*/*pks39*, *pks40*/*pks41* and *pks42*/*pks43*), which are not only adjacent on the chromosome, but which are very closely related in sequence as shown by the phylogenetic trees (Fig. 2). There are also clusters of three (*pks19*/*pks20*/*pks21*), four (*pks29*/*pks30*/*pks31*/*pks32*) and five (*pks5*/*pks6*/*pks7*/*pks8*/*pks9*) PKS genes; in these cases, the KS and AT sequences of members of the clusters are always very closely related. Although some clusters contain genes of similar structure (e.g. *pks22*/*pks23*) there are also cases where the domain structure is different (e.g. *pks24* lacks an MT domain that is present in *pks25*). There are also 13 genes that do not appear to be clustered and these do not have any other PKS genes that are very closely related to them (Fig. 2). Four of the probable pseudogenes occur in clusters. The cluster *pks11*/*pks12* consists of two pseudogenes. As *pks43* seems to be a pseudogene, the cluster *pks42*/*pks43* would contain only one functional gene. *pks20* seems to be a pseudogene, but *pks19*/*pks21* could still form a functional cluster.

Most of the introns in the PKS genes occur in the KS domains. In order to compare the positions of introns in different genes, the deduced protein sequences were aligned to

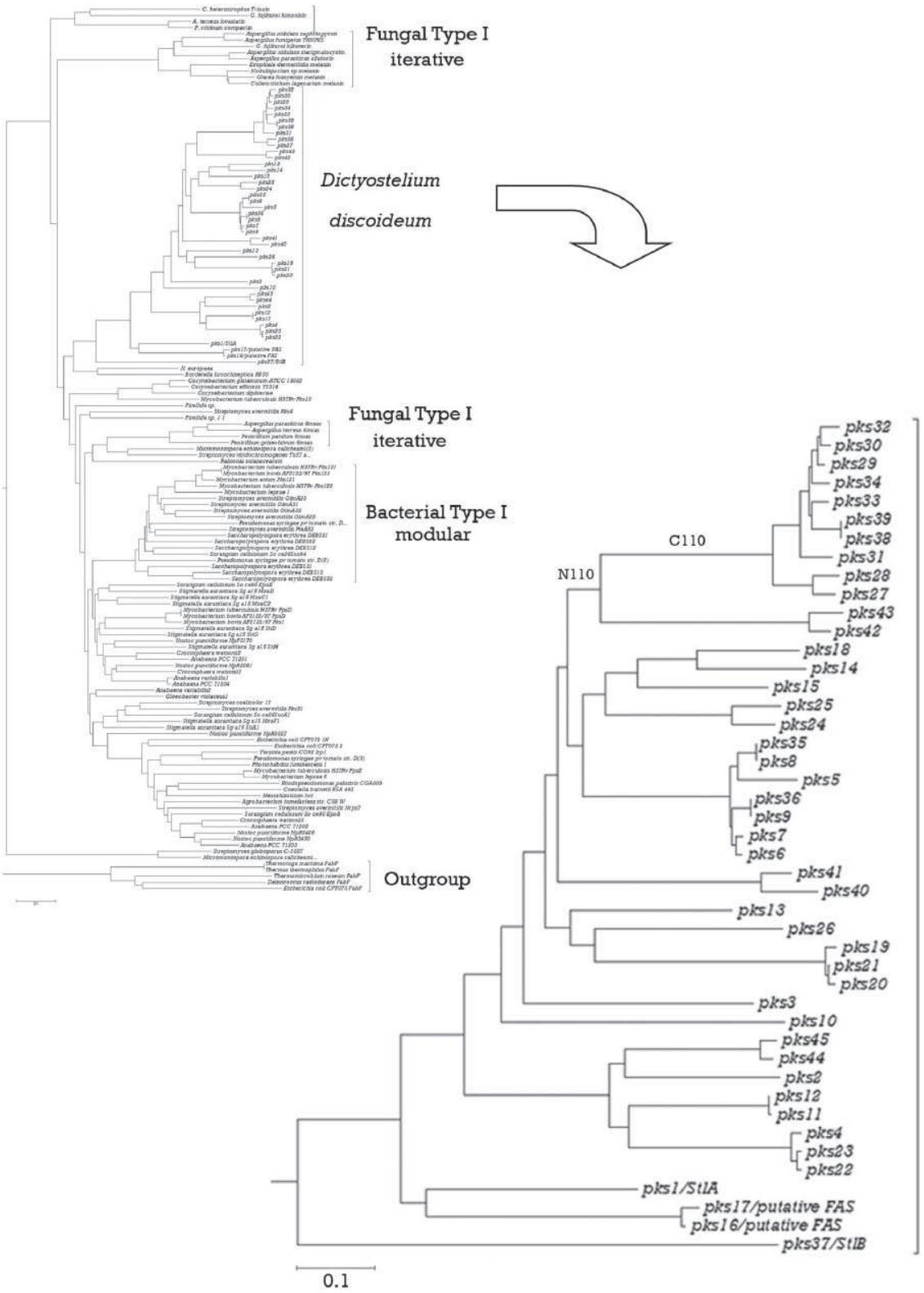

Fig. 2. Phylogenetic tree of the amino acid sequences of KS domains from bacterial, fungal and *D.discoideum* PKSs. The *D.discoideum* sequences form a distinct clade (bootstrap value 92%) which is expanded in this figure. Clades carrying the introns N110 and C100 are indicated. The scale is percentage of amino acid distance.

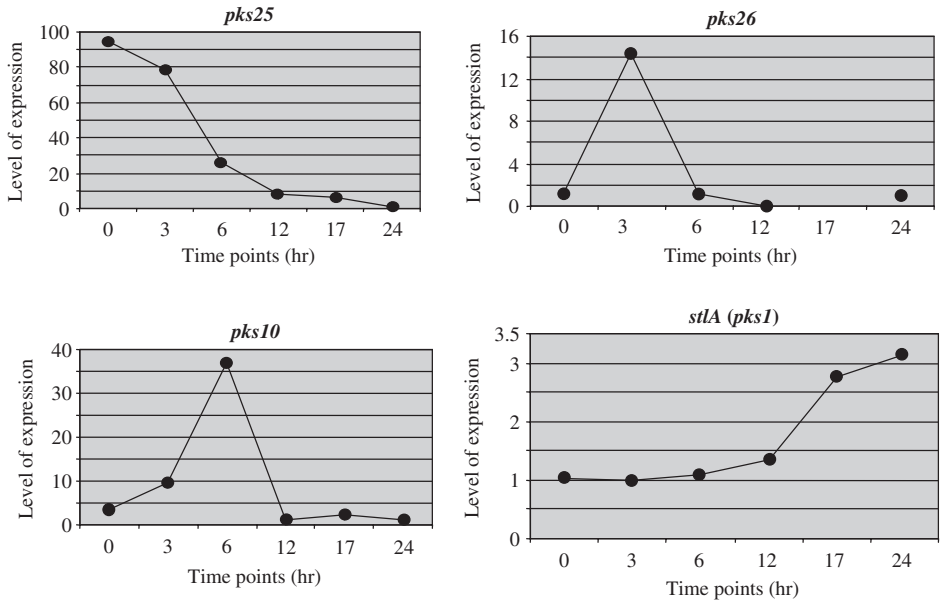

Fig. 3. Quantitative time course analysis of the gene expression of four PKS genes using real time PCR. The transcript levels were normalized using the constitutively expressed gene *IG7*.

Pfam HMM-profiles for the N-terminal (accession number PF00109.16) and C-terminal (accession number PF02801.12) parts of KS. The amino acid number of the profile at the intron position was used to localize the intron. Four intron positions were each found in 10 or more genes. The 10 PKS genes with an intron at position C100 form a clade (Fig. 2; bootstrap value 99%); the intron at position N205 shows an identical distribution except that an intron is also present at this position in *pks2*. The intron at amino acid positions N110 was present in all genes which contain C100 as well as *pks42* and *pks43*; these genes are also a clade (bootstrap value 47%). The intron C248 is present in 34 of the PKS genes. Eight of the genes which lack this intron are more distant in the KS tree (*stlA*, *stlB*, *pks16*, *pks17*, *pks11*, *pks12*, *pks44* and *pks45*). However, it is striking that it is also missing from three genes, whose neighbours all possess the intron (*pks10*, *pks15* and *pks19*), which suggests that the intron has been lost in these cases.

3.3 The transcription of PKS genes

Examination of expressed sequence tags in dictyBase version 2.5 indicated that most of the PKS genes (32/45) were transcribed. Two of the probable pseudogenes (*pks4* and *pks20*) have ESTs. Although the presence of EST clones proves transcription of the genes, it gives little information about timing and level of expression. A number of PKS gene sequences exist on standard microarray chips (G.Shaulsky, personal communication). These data, which can be found in the dictyBase version 2.5, indicate that the genes are transcribed. However, the high similarity of some *Dictyostelium* PKS genes would cause extensive cross hybridization. As a result, there is little correlation of expression patterns with more discriminating quantitative RT–PCR experiments.

The kinetics of expression of 9 PKS genes were followed using real time PCR. The results were normalized to the transcript of a constitutive gene, *IG7* (Huang *et al*., 2006). Starvation was initiated at the start of the time course. After ~6 h cells begin to aggregate and spores and fruiting bodies are formed after ~24 h. Expression patterns could be divided into four categories. Gene *pks18* was expressed during growth but mRNA disappeared within 3 h of starvation. Genes representing a second transcriptional class (*pks24*, *pks25*, *pks26*) were expressed during early development after 3 h of starvation, but their expression had fallen by 6 h. Within this class there were detailed differences in kinetics (e.g. *pks25*, *pks26*; Fig. 3). Levels of gene expression for a third transcriptional class (*pks2*, *pks3*, *pks10*, *stlB* = *pks37*), peaked after ~6 h (Fig. 3) as the cells began to aggregate. A final category represented by a single example (*pks1* = *stlA*) was transcribed preferentially in late development, up to the formation of spores and fruiting bodies after 24 h (Fig. 3). The absolute levels of transcription varied with the measured peak levels differing ~80-fold between the highly expressed gene *pks3* and the weakly expressed gene *stlA*. These results suggest that the PKS genes examined were differentially expressed during different growth stages of *D.discoideum*.

The differential expression of the different PKS genes ought to be reflected in their promoter sequences. In order to examine the probable promoters, a sequence of 1 kb upstream of the start codon was examined for each of the 45 PKS genes. Initial analysis showed that, although the sequences of the coding regions of the PKS genes are highly conserved and can easily be aligned with each other, the similarity usually breaks off upstream of the start codon. Little is known of promoter structure in *D.discoideum*, which are very AT-rich (>95%), so the upstream sequences of the PKS genes were analysed to

identify motifs corresponding to possible binding sequences for transcription factors. The MEME program (Bailey and Elkan, 1994) was used to find 800 motifs of 6–17 bp long. After filtering out very similar motifs, further analysis was carried out using the remaining 676 motifs. The number of conserved motifs was calculated for each pair of PKS genes and most pairs showed little similarity. However, five pairs of genes had very similar upstream sequences (*pks8*/*pks35*, *pks4*/*pks23*, *pks9*/*pks36*, *pks38*/*pks39* and *pks30*/*pks32*). Only the last two pairs belong to clusters and the members of other clusters do not show many conserved motifs. However, a single conserved binding site would be enough to allow coordinated expression. The results with the PKS upstream regions were compared to a set of genes that would be expected to have conserved promoter regions (29 actin genes) and to a set of 47 randomly chosen *D.discoideum* genes. The 1 kb upstream regions of these genes were analysed to identify possible conserved motifs. On average, conserved motifs occurred 33 times in each actin gene upstream region compared to 3 times for the random genes. For the PKS genes the motifs occurred 12 times on average so that the upstream regions were much more similar to each other than those of the random genes. A similar picture was obtained with the numbers of common motifs between pairs of promoters: for the actin gene pairs common motifs occurred 10.6 times on average, compared to 0.6 times for the random genes and 2.6 times for the PKS genes.

The MAST program (Bailey and Elkan, 1994) was used to look for the PKS motifs in the upstream regions of all identified (14102) *D.discoideum* genes. This was used to identify motifs and combinations of 2 or 3 motifs that were specific for PKS genes (present in at least 5 PKS genes and present in less than 30 non-PKS genes). This identified 6 single motifs, 16 motif pairs and 53 motif triples containing 28 different motifs. None of the six single motifs showed any similarity to known regulatory sequences in the TRANSFAC database (Wingender *et al*., 1996). The motifs present in each PKS promoter region and the matches with the TRANSFAC database are shown in the Supplementary Material. The program AlignACE, which is based on a Gibbs sampling procedure (Hughes *et al*., 2000; Roth *et al*., 1998), was also used to find motifs. The PKS promoter regions were analysed using a range of parameter values to mimic the parameter space considered by the MEME program. The 28 PKS-specific motifs identified with MEME were compared with those of the same length found with AlignACE by calculating the average Pearson correlation between base frequencies for all positions in the two motifs. The locations of the AlignACE motif with the highest correlation score were compared with those of the corresponding MEME motif. Twenty of the motifs showed the same locations in the promoter regions (median distance between the MEME and AlignACE motifs not more than 2 bp). Thus, most of the MEME motifs were also found by AlignACE. A comparison of the MEME and AlignACE motifs is shown in Supplementary Material.

4 DISCUSSION

Forty-five PKS genes (including five probable pseudogenes) were identified distributed between the six *D.discoideum* chromosomes. In comparison to the genome annotation in dictyBase version 2.5, 17 PKS genes were added and 20 suggested PKS genes were removed. It is not surprising that an automatic annotation system is relatively inefficient in recognizing genes encoding complex multi-domain proteins when introns are present. The use of HMM-profiles and alignment with known genes helped to identify probable introns. All the genes had a structure typical of Type I iterative PKS genes. In addition to known domains, there were substantial protein coding regions between the AT and ER domains and sometimes after the ACP domain. These may be additional domains of unknown function. Determining the activities of new domains will require an investigation of the chemical structures of the polyketide products as no similarities to known proteins were found.

Phylogenetic analysis of the protein sequences showed that the *D.discoideum* PKS genes formed a discrete group separate from fungal and bacterial sequences (Fig. 2). The chalcone-like genes (*stlA* and *stlB*) and the probable FAS genes (*pks16* and *pks17*; Eichinger *et al*., 2005), were more distant from the others. A very unusual observation is that the genes occur in clusters of 2, 3, 4 or 5 very similar genes. This suggests that the clusters arose from duplications after evolutionary separation from the other PKS genes. In other organisms, PKS genes are usually present in single copies and the proteins are probably homodimers. Although there are some fungal systems (e.g. lovostatin) in which there are two PKSs involved, these PKSs have distinct activity and differ a lot in sequence (Schumann and Hertweck, 2006), unlike the case in *D.discoideum*. It is tempting to speculate that gene pairs allow the formation of heterodimers, which could perhaps extend the biosynthetic repertoire of Type I iterative genes by dividing successive synthesis steps between the two polypeptides. However, it is not clear why this would occur so often in *D.discoideum* and not be observed in other species. It will be interesting to see if this pattern is repeated in other related organisms. This unusual feature of the PKS genes will make it very interesting to characterize their gene products and to see if all the genes in a cluster are needed for successful biosynthesis.

Many of the genes had ESTs in the database. Thus, most of the genes are transcribed and it is likely that most are also translated. Two of the probable pseudogenes also have EST clones. It is conceivable that *trans* complementation between domains occurs as has been reported in other PKSs (Simunovic *et al*., 2006). Each of the nine PKS genes tested showed characteristic kinetics of expression. Most of them were induced by starvation, which supports the idea that the PKS genes may be involved in activities such as signalling between cells or protecting the differentiating organism from competitors and predators. In contrast to strong conservation of the protein sequences, the regions upstream of the PKS genes showed little conservation. Potential transcription factor-binding motifs were identified, but more experimental data on expression levels and kinetics are necessary to narrow down the significant motifs.

No genes encoding Type I modular PKSs (Weissman and Leadlay, 2005) or Type II PKSs (Petkovic *et al*., 2006) were identified. Similarly no non-ribosomal peptide synthetase genes were found. The genomes of several other Dictyostelid species

are about to be sequenced, including *D.purpureum*, *Polysphondylium violaceum* and *D.citrinum* (Baylor College of Medicine/Rice University/Joint Genome Institute) and *D.mucoroides* (G.Gloeckner personal communication.). These genomes may reveal yet greater varieties of PKS or non-ribosomal peptide synthetase genes. The potential genetic diversity of natural products in Dictyostelid populations, which are ubiquitous in forest and cultivated soils, is very high. We infer from the fact that these large genes have maintained their ORFs in the face of genetic drift that they are critical to the survival of the organisms whether, as in the case of DIF, they are used as developmental signalling molecules or to control potentially harmful bacteria, fungi or nematodes.

ACKNOWLEDGEMENTS

We would like to thank Gad Shaulsky and Jessica Svetz of the Baylor College of Medicine for help with the real time PCR experiments. The work of the curators of the dictyBase, Pascale Gaudet, Karen Pilcher and Petra Fey was critical to the annotation of the PKS genes. This work was supported by grant TP-05/0058-23 (to D.H.) from the Ministry of Science, Education and Sports, Republic of Croatia, 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.) and by a stipendium of the DAAD (to J.Z.), by a grant from the Slovenian Research Agency (to B.Z. and T.C) and by a grant from Ad-Futura, Science and Education Foundation of the Republic of Slovenia, Public Fund (to T.C.).

Conflict of Interest: none declared.

REFERENCES

Akaishi,E. *et al.* (2004) Differentiation-inducing factor-1-induced growth arrest of K562 leukemia cells involves the reduction of ERK1/2 activity. *Eur. J. Pharmacol.*, **485**, 21–29.

Altschul,S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Ansari,M.Z. *et al.* (2004) NRPS-PKS: a knowledge-based resource for analysis of NRPS/PKS megasynthases. *Nucleic Acids Res.*, **32**, 405–413.

Austin,M.B. *et al.* (2006) Biosynthesis of *Dictyostelium discoideum* differentiation inducing factor by a hybrid type I fatty acid–type III polyketide synthase. *Nat. Chem. Biol.*, **2**, 494–502.

Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **2**, 28–36.

Bateman,A. *et al.* (2002) The Pfam protein families database. *Nucleic Acids Res.*, **30**, 276–280.

Chisholm,R.L. *et al.* (2006) dictyBase, the model organism database for *Dictyostelium discoideum*. *Nucleic Acids Res.*, **34**, D423–D427.

Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.

Eichinger,L. *et al.* (2005) The genome of the social amoeba *Dictyostelium discoideum*. *Nature*, **435**, 43–57.

Haydock,S.F. *et al.* (1995) Divergent sequence motifs correlated with the substrate specificity of (methyl)malonyl-CoA:acyl carrier protein transacylase domains in modular polyketide synthases. *FEBS Lett.*, **374**, 246–248.

Hranueli,D. *et al.* (2005) Plasticity of the *Streptomyces* genome – evolution and engineering of new antibiotics. *Curr. Med. Chem.*, **12**, 1697–1704.

Hughes,J.D. *et al.* (2000) Computational identification of *cis*-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J. Mol. Biol.*, **296**, 1205–1214.

Huang,E *et al.* (2006) bZIP transcription factor interactions regulate DIF responses in *Dictyostelium*. *Development*, **133**, 449–458.

Jenke-Kodama,H. *et al.* (2005) Evolutionary implications of bacterial polyketide synthases. *Mol. Biol. Evol.*, **22**, 2027–2039.

Kessin,R.H. (2001) *Dictyostelium – Evolution, Cell Biology, and the Development of Multicellularity*. Cambridge University Press, Cambridge, pp. 1–294.

Khosla,C. *et al.* (2007) Structure and mechanism of the 6-deoxyerythronolide B synthase. *Annu. Rev. Biochem.*, **76**, 195–221.

Kubohara,Y. *et al.* (2003) DIF-1, an anti-tumor substance found in *Dictyostelium discoideum*, inhibits progesterone-induced oocyte maturation in *Xenopus laevis*. *Eur. J. Pharmacol.*, **460**, 93–98.

Kumar,S. *et al.* (2004) MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. *Brief. Bioinformatics*, **5**, 150–163.

Petkovic,H. *et al.* (2006) Genetics of *Streptomyces rimosus*, the oxytetracycline producer. *Microbiol. Mol. Biol. Rev.*, **70**, 704–728.

Rice,P. *et al.* (2000) EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet.*, **16**, 276–277.

Roth,F.P. *et al.* (1998) Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat. Biotechnol.*, **16**, 939–945.

Rutherford,K. *et al.* (2000) Artemis: sequence visualisation and annotation. *Bioinformatics*, **16**, 944–945.

Schneider,T.D. and Stephens,R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, **18**, 6097–6100.

Schumann,J. and Hertweck,C. (2006) Advances in cloning, functional analysis and heterologous expression of fungal polyketide synthase genes. *J. Biotechnol.*, **124**, 690–703.

Schweizer,E. and Hofmann,J. (2004) Microbial Type I fatty acid synthases (FAS): major players in a network of cellular FAS systems. *Microbiol. Mol. Biol. Rev.*, **68**, 501–517.

Shaulsky,G. and Loomis,W.F. (1995) Mitochondrial DNA replication but no nuclear DNA replication during development of *Dictyostelium*. *Proc. Natl Acad. Sci. USA*, **92**, 5660–5663.

Simunovic,V. *et al.* (2006) Myxovirescin A biosynthesis is directed by hybrid polyketide synthases/nonribosomal peptide synthetase, 3-hydroxy-3-methyl-glutaryl-CoA synthases, and trans-acting acyltransferases. *Chem. Biochem.*, **7**, 1206–1220.

Thompson,J.D. *et al.* (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.

Van Driessche,N. *et al.* (2005) Epistasis analysis with global transcriptional phenotypes. *Nat. Genet.*, **37**, 471–477.

Weissman,K.J. and Leadlay,P.F. (2005) Combinatorial biosynthesis of reduced polyketides. *Nat. Rev. Microbiol.*, **3**, 925–936.

Wingender,E. *et al.* (1996) TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res.*, **24**, 238–241.

Discussion

3. Discussion

Through time humankind was struggling for survival and continuation of the species by supplementing its physical deficiencies using its "superior" knowledge and intelligence. Today humankind is still facing various hazards most of which are a direct result of our own "ingenuity" such as unsustainable population growth, overexploitation of natural resources and pollution. During the last 30 years, with the appearance of bacteria resistant to antibiotics, we are also facing medical challenges. Today we are faced with a shortage of antibiotic drugs as currently only a handful of antibiotics can still be effectively used in human medicine. The causes for developing resistance in bacterial species to popular antibiotics are mostly caused by human misdoings like unnecessary and uncontrolled usage by humans, usage as animal feeds, in agriculture, etc. The other reason lies in the pull-out of big pharmaceutical industries from this field of research, due to its high risk and low profit return, which is now mostly done by smaller pharmaceutical companies and various University groups (Demaine, 2009). Natural products are major players in the field of antibiotic drugs with non ribosomal peptides and polyketides making up majority of them. Traditional methods for discovery of novel natural products involved their purification and elucidation from microbial fermentations. With advances in understanding mechanisms of biosynthesis and availability of microbial genomes new approach - genome mining arises (Lautru *et al*., 2005; Zucko *et al*., 2010).

To speed-up and make the"hunt" for modular biosynthetic clusters easier we have developed a program package called *ClustScan*. In it a top-down genome mining approach was implemented to search the nucleotide sequences, having either genomic or metagenomic origins, for modular polyketide synthase (PKS) and non ribosomal synthetase (NRPS) biosynthetic clusters. When a cluster is identified, based on gene prediction implemented using Glimmer (Delcher *et al*., 2007) and/or Genemark (Besemer and Borodovsky, 2005) programs, the similarity search executed using HMMER (Eddy, 1998), and cluster organisation and biosynthetic order reconstructed by the user, the program is able to predict the chemical compound synthesised by the modular biosynthetic cluster. To be able to generate a chemical product of the biosynthetic cluster all components (domains) of the

system with all their properties (activity, specificity) have to be determined and hierarchically organised. Until now the program is able to predict chemical structure only for modular PKS gene clusters as the specificity for NRPS adenylation (A) domains in not implemented yet due to large number of amino acids (groups) used as substrates. However, this process is currently under way. The domain identification is obtained using HMM profiles for respective domains found in PKSs and NRPSs using HMMER (Eddy, 1998) as a search tool. HMM profiles were built using domains from curated complete bacterial PKS and NRPS clusters. This process is independent from the gene finding results as it is done directly on all 6-reading frame translation of genomic sequence, thus increasing probability of finding all genes of interest which might be skipped if only ORFs recognised by gene finding software would be used for function determination (similarity search). This approach was also tested on annotation of PKS genes in *D. discoideum*, which contain introns.

Compared with the published annotation of the *D. discoideum* genome project (http://dictybase.org) the tested method was shown to be superior in identifying PKS genes and had made significant changes to more than three quarters of previously identified PKS genes. Initially HMM profiles used in *Dictyostelium* PKS annotation were built from bacterial type I modular PKSs but to increase the sensitivity of the profile they were refined by curated PKS domains from *D. discoideum*. Using a combination of local and global HMM profiles was sensitive enough not just to define the position of PKS domains within the genome but also to locate almost exact location (position) of introns within the domains (Fig. 5). The question arises why Pfam (Finn *et al*., 2008) profiles were not used for similarity search when they were already available for majority of PKS domains? There were several reasons why custom-made profiles were used. One lies in the sensitivity of custom made profiles. Although Pfam profiles are built using proteins from wide range of organisms they do not deal well with organism specific deviations (exceptions), and *D. discoideum* is also evolutionary distant from the majority of organisms (Bapteste *et al*., 2002) which might cause problems with sensitivity of the profiles. Pfam profiles do not always cover the whole length of the domain. Such an example was the keto synthase (KS) domain whose Pfam profile was more than 50 amino acids (more than 10% of total length of the domain) shorter than KS domains from the literature (Kroken *et al*., 2005, Jenke-Kodama *et al*., 2005). The last reason is that HMM profiles were intended only for annotation of polyketide genes in *D.*

discoideum, and not as a general profiles, so it was logical to make them as sensitive as possible for the selected organism.

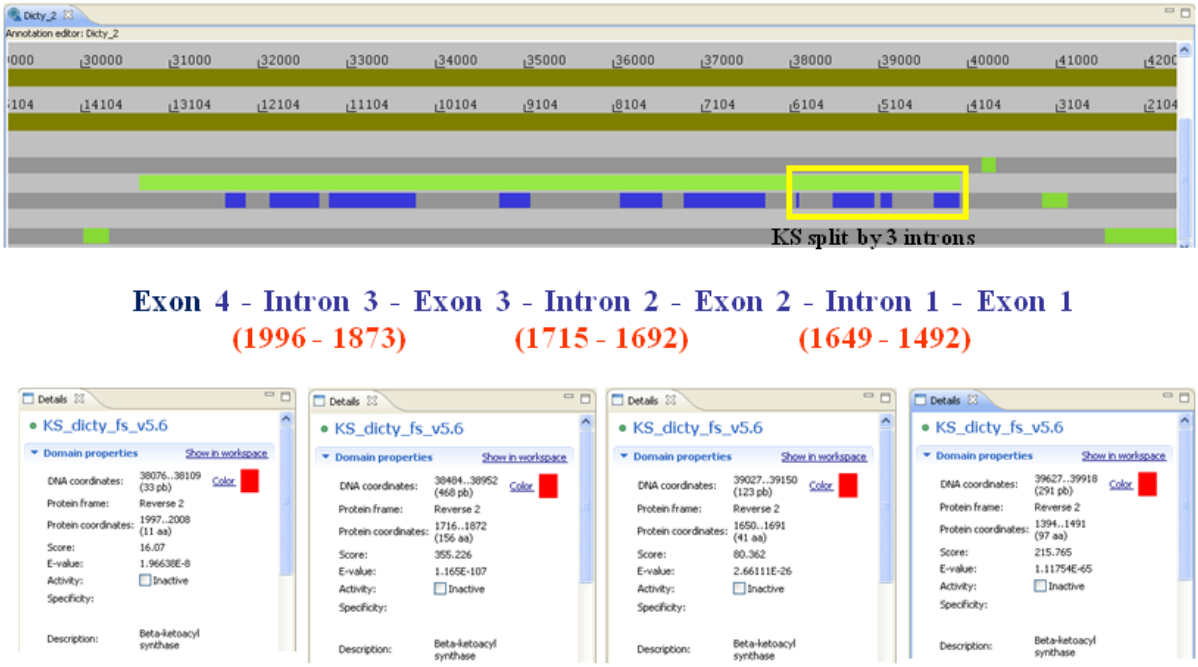

Fig. 5. *ClustScan* window showing analysis of the *D. discoideum pks29* gene. ORFs found by Glimmer are shown as green bars while PKS domains found by HMMER are represented as blue bars. Highlighted with yellow box is a KS domain consisting of four exons split by three introns. Shown below are four Details windows with the results of similarity searches obtained by HMMER using local HMM profile for the KS domain. From these data it is possible to reconstruct the entire KS domain – a principle used in annotation of the *D. discoideum* PKS genes. After the ACP domain (first HMMER hit from the left), there is a region without known function, typical for most of PKS genes found in *D. discoideum*.

The annotation process of *D. discoideum* discovered interesting results concerning the structural organisation of its PKS genes – large regions of the gene did not show similarity to known proteins. From the two regions between AT and ER domains one upstream from ER appears to be a structural KR domain (Keatinge-Clay and Stroud, 2006). This domain appears to be conserved structurally while on the sequence level is quite diverged, which is a reason why it was not found before the crystal structure was solved. The region downstream of AT domain as well as the region after the ACP in some of the genes still has not been functionally identified. The region downstream of the AT domain is present in the majority of PKS genes and may play a role in structural positioning of the entire complex. Until now

Dictyostelium is the only organism harbouring PKS genes that have substantial region after the ACP domain without known function. Purely speculatively, two scenarios explaining this might be possible - this is either a previously uncharacterised domain specific for *Dictyostelium* PKS genes or that these regions might have been chalcon synthases, still present in *stl*A and *stl*B genes, which have lost their function during the evolution and were degraded over time.

Phylogenetic analysis of the KS domains and whole PKS genes positions *Dictyostelium* in a clade distinct from fungal and bacterial sequences. Although the phylogenetic tree does not support relatedness with fungal PKS, the organisation of genes tends to put *D. discoideum* PKS genes closer to fungal than bacterial PKSs. One reason lies in biosynthetic gene organisation - type I fungal PKSs are iterative, while bacterial type I PKSs are modular. Some *D. discoideum* PKS genes appear clustered on a chromosome, resembling modular PKSs, but it appears that clustering observed at a DNA level is not transferred into the biosynthetic pathway. Up to now it was proven that the PKS/chalcone hybrid genes (*stl*A and *stl*B) follow an iterative biosynthetic logic (Austin *et al*., 2006; Ghosh *et al*., 2008), and the same is assumed for the remaining PKS genes. The other reason is the occurrence of methyl transferase (MT) domains. This domain rarely appears in type I modular bacterial PKS while it is quite common in type I iterative fungal PKSs - with almost 80% of PKS genes harbouring a MT domain. Clustering of PKS genes on a chromosome and high sequence similarity of, usually neighbouring, PKS genes would support the theory of PKS gene evolution by gene duplication (Jenke-Kodama *et al*., 2005; Ridley *et al*., 2008). Also interesting is the occurrence of clade-specific introns which favours the idea of gene duplication rather than the possibility that introns were introduced by independent events.

In the annotation of *D. discoideum*, a manual annotation approach was used as we only aimed on annotating one gene family in a single organism. The next logical step was automation of the process which was implemented in the *ClustScan* program package. As *ClustScan* is intended to be used for annotation of modular biosynthetic enzymes from various organisms profile HMMs used for domain identification were not refined with domains from a single organism, like the ones used for *D. discoideum* annotation. Instead domains from curated complete modular PKS/NRPS gene clusters spanning through various

bacterial families were used for building HMM profiles. Such an approach increased the sensitivity of HMM profiles, allowing detection of remote homologues and domain families with more divergent members. This was shown to be more successful compared with methods relying on BLAST [SEARCHPKS (Yadav *et al*., 2003), ASMPKS (Tae *et al*., 2007)] for identification of PKS domains, especially for DH and ACP domains.

When domains are recognised using HMM profiles the next step is to determine activity or specificity, which is until now implemented only for PKS domains. KS and ACP domain are, if found, considered to be active, as they form a minimal PKS module. One more domain forming the minimal PKS module – AT is assumed to be active if present but is used to determine its specificity. A method similar to Minowa and collaborators (Minowa *et al*. 2007) was tested as one possibility to determine specificity. Although the method was able to discriminate between various substrate specificities of AT domains statistical parameters returned by HMMER were inadequate to be used in an automated computer-based decision making. Instead a method giving simple **yes** or **no** answers based on comparison of extracted motif assembled from specificity determining residues of query domain with libraries of motifs characteristic for specific substrate was used. A similar approach was also used for the KR domain while for the DH and ER domains, a method based on the alignment score of the domain with an HMM profile was used. The reason for this approach was insufficient structural information for PKS reduction domains, except for KR, which would allow prediction of activity. In the case of the DH domain the other problem, which will remain when structural information is available, is the high divergence within the protein family, thus, making the alignment and extraction of activity-determining residues unreliable. For the ER domain there were no data about amino acids affecting its activity and a HMMER scores alone were not enough to discriminate between active and inactive domains, so a combination of score and module structural/organisation information were used. As ER is the last domain in the biosynthetic pathway it requires presence and activity of preceding reduction domains. This property was used to achieve correct activity prediction of the ER domain in 95% of cases in the test sample. In the case of *ClustScan,* the combination of methods for determining activity/specificity of domains based on motif extraction, similarity search statistics and rules following the biosynthetic logic was the best solution which at the end was able to achieve correct predictions in a range from 80 to 95% for all domains.

Compared with other solutions (programs) for annotation and analysis of PKS/NRPS clusters *ClustScan* was shown to be superior in detecting PKS domains due to use of more sensitive similarity search method (HMM profiles) and filtering of results using two cut-offs (srict and relaxed) based on statistical parameters returned by HMMER. Another major advantage of *ClustScan* is the detailed implementation of recent advances in deciphering function and specificities of components making up biosynthethic clusters which enabled it to mimic biosynthetic pathway *in silico* and to predict the chemical structure (compound) synthesised with high accuracy. Comparing with published annotations of tested clusters *ClustScan* gives predictions almost identical to experimentally determined structures. The majority of differences arise due to wrong prediction of the activity of reduction domains and the inability to predict the stereochemistry of hydroxyl and/or methyl groups. In a few cases, wrong predictions for substrate specificity of AT domains are given. This mainly affected AT domains incorporating rare substrates (e.g. methoxymalonyl-CoA, ethylmalonyl-CoA) for which data were scarce, with roughly only 10 domains for each substrate. With an increased number of ATs specific for those substrates, the motifs used for prediction can be modified to better reflect observed pattern in a larger number of clusters, thus increasing its accuracy.

The *ClustScan* program package was also tested on a metagenomic dataset where it identified a fairly large PKS/NRPS hybrid cluster, whose product was not predictable due to the presence of NRPS modules. The analysis of metagenomic sequences also showed dependence of analysis results on the quality of input DNA sequence. As modular biosynthetic clusters are large, with clusters larger than 100 000 nucleotides being not uncommon, to be able to make relevant predictions, the input DNA sequence must contain the entire cluster which must be of high quality, in terms of sequence. In the metagenome data set (Rusch *et al*., 2007) which was used for screening only 119 successfully assembled contigs larger than 100 000 nucleotides were present from more than 7 million sequences deposited in the dataset.

To be able to detect specificity determining residues (SDR) we have been involved in developing of an algorithm that would be able to perform clustering of protein sequences into two or more functional subgroups within a family. The method starts from a multiple alignment (MSA) of the family and ranks columns of the MSA according to the “strength of

the signal" for functional split. The probabilistic model used for ranking columns is a rough approximation of a process describing protein evolution based on a BLOSUM matrix. It gives each column a value which describes how good it models clustering around two or more amino acids rather than a single amino acid (Fig. 6). When ranking is complete, the user selects a motif from the top ranked columns. The exact number of relevant positions varies from family to family, so the user is able to define the number of positions to be used for motif construction. However, the length of the motif used does not affect strongly the accuracy of clustering, as was shown with varying length of the motif, but rather the strength of the signal encoded within each column of the MSA. The motif is then used for clustering of the family into subtypes.

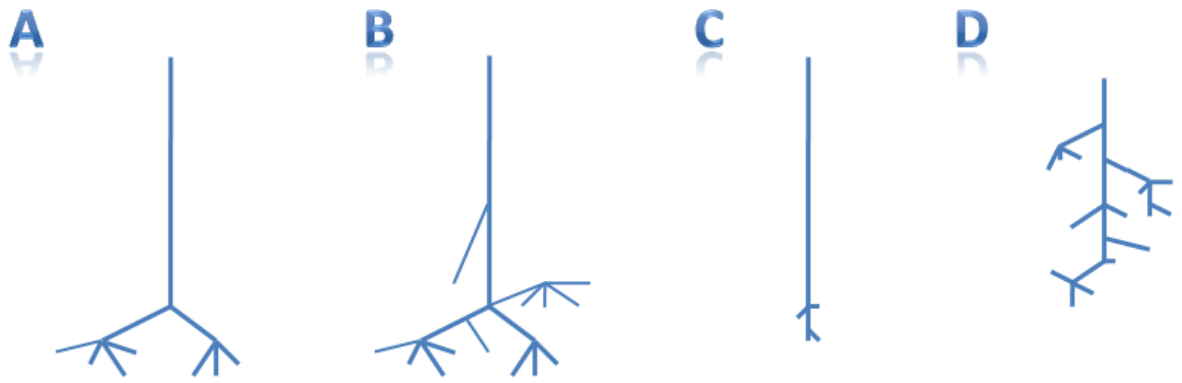

Fig. 6: Models describing possible evolutionary process of amino acids within one column of the multiple alignment. Amino acids are represented as lines with distance corresponding to distance described by BLOSUM matrix. A: model describing ancestral split into two clusters of related amino acids – observed in clustering of malyonyl/methylmalonyl specific AT domains among top ranked positions, B: example of three-way split with "evolutionary noise" from unrelated amino acids. C: pattern seen in conserved columns usually corresponding to active site or structurally important residues, D: pattern common for column not under selective pressure

Since the only input for the method is MSA of the protein family it is important to emphasise that the quality of the analysis depends on the quality of given alignment. If there are inaccuracies or misalignments, especially in the core regions for determining specificity, the alignment should be remade from either only conserved regions within the protein family or with a reduced subset of "non problematic" sequences. Such an example of a problematic multiple alignment was the protein kinase family used in testing of the method. When all

members of the serine/threonine and tyrosine subfamilies were extracted from a reference database (http://pkr.genomics.purdue.edu/pkr/), the multiple alignment created contained only a few columns without gaps, which could be used by the program. The resulting analysis was unable to determine any meaningful clustering from the given data. After highly diverged and short sequences causing the problem have been removed, the analysis clustered members of the family with 100 % accuracy into respective subfamilies.

The results obtained by the method were promising as it clustered the test protein families into biologically relevant subgroups based only on their respective multiple alignments. Also it is important that the length of the motif used for clustering does not have a strong influence on the accuracy of clustering for the majority of families. This simplifies the automation of the method as it is possible to define a default length of the motif to be used for clustering without strongly effecting accuracy. The protein family which showed the highest dependence of clustering accuracy on motif length was the PKS ketoreductase domain. The reasons for this discrepancy might lie in the functional duality of ketoreductase domain reflected in controlling activity and stereochemistry of ketoreduction as well as stereochemistry of methyl groups. This multifunctionality is also reflected at a sequence level with higher divergence within columns of the multiple alignment (more evolutionary noise) thought to be important for determining subtype than in other tested families. The other tested family of special interest was the PKS AT domain for which prediction of substrate specificity was implemented in *ClustScan*. As a test sample, the two largest subgroups were used – AT domains specific for malony-CoA and methylmalonyl-CoA. Previously published positions (Yadav *et al*., 2003) involved in substrate specificity were also identified by our method and were among 30 highest ranked positions. Depending on the length of the motif used for clustering, the error was between 1 % and 3 % which is slightly better than predictions implemented in other software for analysis of modular PKS. Testing with more subgroups (3) has not yielded satisfactory results with more than third of test sample falling into the wrong cluster. This is not unexpected as members of each subtype were, due to limited number of AT domains specific for those substrates with specificity experimentally determined, not well balanced – with two of the groups being underrepresented. In order for the method to be able to reliably calculate significant positions in the sample it has to contain roughly 30 members of each subtype (P. Goldstein

personal communication). Another logical possibility for such a high number of wrongly clustered domains might be that residues determining the specificity of AT domains incorporating larger substrates are not the same as ones determining the functional split for malonyl/methylmalonyl subgroups. This assumption is strengthened by the residues indicative for methoxymalonyl-specific AT domains. Motif used in *ClustScan* for identification of methoxymalonyl incorporating AT domains is not the same as the motif used to identify specificity of malonyl/methylmalonyl AT domains. To add further confusion, phylogenetic analysis of AT domains shows relatedness with methylmalonyl specific AT domains as they form a joint clade, separate from malonyl specific AT domains (Yadav *et al*., 20003; Jenke-Kodama *et al*., 2005). The underlying principle for evolutionary split statistics is based on BLOSUM 50 matrix and it is possible that it is inadequate for this specific case where sequences are highly similar and evolutionary closely related.

While some residues are important in discriminating between subtypes of protein domains other residues play a role in interaction between domains. Interactions within proteins forming the type II PKS complex responsible for biosynthesis of daunorubicin and doxorubicin in *S. peucetius* (*Grimm et al*., *1994*) were investigated using protein-protein docking simulations. The purpose of docking simulations of type II PKS enzymes was to compare with experimental data obtained using the yeast two-hybrid system about interactions of PKS complex subunits, as well as to obtain, if possible, information about structural organisation of entire complex. As none of the proteins constituting doxorubicin PKS had their crystal structure solved they had to be modelled using computational tools to enable us to carry out docking simulation. As all of the proteins had orthologs with already solved structures homology modelling was chosen to obtain near-native, biologically relevant structures. To minimise method errors two programs for homology modelling were used – MODELLER (Eswar *et al*., 2007) and Swiss-Model (Arnold *et al*., 2006). Models obtained from both methods were evaluated and ones with more favourable energy profiles were used for docking simulations. To rule out uncertainty and unreliability associated with blind docking methods two pipelines for protein – protein docking were used – PatchDock (Duhovny *et al*., 2002) and ClusPro (Comeau *et al*., 2004). These methods were chosen as both are rigid body docking algorithms and they both incorporate methods for filtering and clustering of docked solutions which significantly reduces the number of relevant docking

solutions, thus enabling manual comparison. Docking scenarios from both methods were compared and if the RMSD was smaller than 4 Å it was treated as correct docking scenario.

Results of protein docking confirmed most of the interactions observed in the yeast two-hybrid system. Those involved interactions of KS subunits, cyclase and ketoreductase subunits. Interactions not confirmed by computational methods are those involving the ACP domain. Although the ACP domain is a part of minimal PKS none of the simulations returned docking solutions common for both methods. It has been hypothesised that ACP forms weak and transient interactions with other members of the minimal PKS (Hertweck *et al*., 2007) and such interactions would probably not be detectable using docking algorithms. The other reason why interactions involving this domain were not observed with computational docking methods might lie in the decision to use a rigid body docking algorithm instead of a flexible docking algorithm. If the protein undergoes significant conformational changes during the interaction, a rigid body docking algorithm won't be able to compensate for it and will fail in finding a correct docking solution. As rigid body algorithms were able to predict correct docking solutions even for proteins making tight complementary interfaces (KSα-KSβ) (Keatinge-Clay *et al*., 2004) it is unlikely that the choice of docking algorithm was wrong. It seems that interactions involving ACP protein are elusive for computational docking approaches and further experiments are necessary.

Future prospects

4. Future prospects

In this thesis PKSs have been used as a test bench for bioinformatic methods ranging from gene finding and identification, specificity determination to protein modelling and docking. The next logical step was to deal with a question how these huge biosynthetic clusters arose and evolved over time. Which processes induced such vast diversity both of synthesised products as of genetic makeup and organisation with retaining the same mechanism of biosynthesis? With routine genome sequencing more cluster sequences are becoming available thus allowing evolutionary studies of these enzymes.

Pioneering work in phylogenetic studies of PKSs suggested that modular PKSs share a complex evolutionary history with iterative PKS and animal and bacterial FAS (Kroken *et al*., 2003; Jenke-Kodama *et al*., 2005). Evolutionary mechanisms that play an important role in evolutionary diversification of modular PKS are believed to be: gene and module duplication, homologous recombination and horizontal gene transfer (Ridley *et al*., 2008). The primary mechanism responsible for evolutionary diversifications of PKSs is believed to be module duplication. These finding are based on phylogenetic analysis of KS domain which showed higher similarity of KS domains within the cluster than with KS domains from other clusters (Jenke-Kodama *et al*., 2006). The AT domains, on the other hand, are clustered in phylogenetic trees based on their substrate specificity. Therefore, AT domains with different substrate specificity are more similar to AT domains from other cluster having the same substrate specificity than to AT domains of the same cluster having different substrate specificity (Yadav *et al*., 2003; Jenke-Kodama *et al*., 2005; Jenke-Kodama *et al*., 2006). Based on phylogenetic analysis of KS and AT domains it was hypothesised that modular PKS clusters evolved by duplication of an ancestral module followed by recombination events that replaced the AT and reduction domains to generate different module specificities (Jenke-Kodama *et al*., 2006).

We have applied another approach. Instead of collecting domains from unrelated (individual) clusters, or clusters from a single organism, we have searched the literature and sequence databases for orthologous clusters. Orthologous clusters were defined based on similarity of cluster sequence, cluster organisation and the synthesised product. In total 6

groups of orthologous clusters containing 17 modular PKSs, were found and later used for phylogenetic analysis. As expected, in many cases, domain pairs from orthologous clusters clustered together as the most similar pairs. However, gene conversion appeared to be common as, in some cases, domains within the cluster were more closely related. This was observed for all the domains but with falling frequency based on domain's position within the cluster.

The other currently undergoing part of research is dealing with the conservation of metabolic pathways within a range of organisms. For that purpose the presence of enzymes constituting shikimic acid pathway was investigated in the predicted proteomes of almost 500 prokaryotes, which were deposited at NCBI's genome database (http://www.ncbi.nlm.nih. gov/sites/genome), using HMM profiles. Enzymes of the shikimic acid pathway play an important role as they are responsible for supplying precursors for aromatic amino acid biosynthesis, as well as for synthesis of other aromatic compounds. This pathway has been found in bacteria and plants (Herrmann, 1995) and recently in certain apicomplexan parasites (Roberts *et al*., 1998), but is not present in animals and they must obtain these compounds from dietary sources. In bacteria, the shikimate pathway serves almost exclusively for synthesis of aromatic amino acids while in higher plants mostly as precursors for secondary metabolites such as pigments, defence compounds, UV protectors etc. (Herrmann, 1995). Preliminary results had confirmed the presence of the complete shikimic acid pathway in most free living bacteria. Most of the host-associated bacteria (both symbiotic and pathogenic), on the other hand, have an incomplete shikimic acid pathway. Although this finding questions the conventional belief that shikimic acid pathway is essential in prokaryotes, it is possible that organisms missing full pathway substitute it by either from parasitized host or from symbiotic relationship as was found in *Nematostella vectensis* (Starcevic *et al*., 2008).

Also continued is the development of the method for clustering of protein domains (Goldstein *et al*., 2009). Currently method calculates the evolutionary split statistic for each column of the multiple alignment and gives higher values when amino acid distribution within the column is better described by an evolutionary model that assumes clustering around two or more amino acids. To give more relevant information about the importance

of each position, a statistical method based on Bonferroni correction (Cabin and Mitchell, 2000) has been implemented and is currently being tested. The method returns for each position the percentage of significance in respect to evolutionary model describing the clustering. Current results show only a slight improvement in clustering accuracy, mostly on protein families which had a higher error rate in previous work, but make selection of positions used for motif building much easier and more uniform.

Abstract

5. Abstract

With the development of new high capacity DNA sequencing techniques the use of a computational approach in biology is gaining even greater importance. In this thesis several of the bioinformatic methods have been employed on one class of enzymes – polyketide synthases (PKSs) - which were used to decode information stored in DNA into its more useful form - a chemical compound synthesised by the enzyme.

To get the information about the chemical compound synthesized by the enzyme, DNA sequences coding for modular biosynthetic clusters first have to be identified. For that purpose a top-down approach relying on Hidden Markov Model (HMM) profiles, describing all type I PKS domains was used. HMM profiles were chosen due to superior sensitivity coming from capturing information from multiple members of the protein family. Another advantage of HMM profiles is their robustness which was demonstrated in annotation of PKS genes in the genome of *Dictyostelium discoideum* where profiles were “retrained” in several steps with organism-specific sequences and were able, at the end, to accurately detect all deviations within the sequence, as was the case with introns occurring within domains.

When all domains constituting PKS are identified their activity and/or specificity has to be determined. In this thesis several methods were used - comparison of motifs consisting of specificity-determining residues, the statistical parameters of similarity search or predefined rules based on existing knowledge, depending on the type of the domain. After all existing components of the system (all PKS domains) were found and their properties (activity/specificity) determined they were organised into a "functioning system" which is able to predict the chemical entity synthesised by the system.

In addition to the information about protein function and specificity/activity, information about the structure of the protein as well as its interactions can also be extracted from the DNA sequence. Structures of polypeptides constituting the daunorubicin PKS were built from their DNA sequences using homology modelling methods. These structures were later used for rigid body docking simulations which revealed interacting partners and revealed some information about the overall structure of the complex.

5.1 Zusammenfassung

Mit der Entwicklung neuer Hochleistungs-DNA-Sequenzier-Technologien gewinnt die computergestützte Informationsverarbeitung in der Biologie immer größere Wichtigkeit. In dieser Doktorarbeit wurden mehrere bioinformatische Methoden auf eine Klasse von Enzymen, die Polyketidsynthasen (PKS) angewandt. Dies ermöglichte es uns, die in der DNA verschlüsselte Information in eine nützlichere Form umzuwandeln - die chemische Verbindung, die vom Enzym synthetisiert wird.

Um diese Information zu erhalten, müssen zunächst die DNA-Sequenzen identifiziert werden, die für modulare Biosynthese-Cluster kodieren. Zu diesem Zweck wurde hier ein top-down-Ansatz mit Hidden Markov Modell (HMM)-Profilen verwendet, die alle Typ I-PKS-Domänen beschreiben. HMM-Profile sind anderen Ähnlichkeitssuchmethoden aufgrund ihrer höheren Empfindlichkeit überlegen, da sie Informationen von vielen Mitgliedern einer Proteinfamilie verwenden. Ein weiterer Vorteil von HMMs ist ihre Robustheit, wie bei der Annotation von PKS-Genen im Genom von *Dictyostelium discoideum* gezeigt wurde, wo Profile mit organismus-spezifischen Sequenzen „trainiert“ wurden und nach mehreren Trainingsschritten im Stande waren, alle Abweichungen innerhalb der Sequenz zu identifizieren, z. B. Introns, die innerhalb der Domäne vorkommen.

Wenn alle Domänen, die zur PKS-Familie gehören, identifiziert sind, muss ihre Aktivität und/oder Spezifität bestimmt werden. In dieser Arbeit wurden je nach Domänentyp verschiedene Methoden angewendet – der Vergleich von Motiven, die aus spezifitätsbestimmenden Aminosäuren bestehen, statistische Parameter der Ähnlichkeitssuche oder vordefinierte Regeln, die auf schon vorhandenem Wissen über diese Domänen beruhen. Nachdem alle vorhandenen Bestandteile des Systems (alle PKS Domänen) gefunden und ihre Eigenschaften (Aktivität/Spezifität) bestimmt waren, wurden sie in ein "Funktionssystem" organisiert, das im Stande ist, das von diesem System synthetisierte Produkt vorauszusagen.

Aus der DNA-Sequenz kann man außer Information über die Funktionalität und die Spezifität/Aktivität des kodierten Proteins auch noch Information über seine Struktur und die Wechselwirkungen mit anderen Proteinen gewinnen. Strukturen von Polypeptiden, die die Daunorubicin-PKS bilden, wurden aus ihren DNA-Sequenzen über Homologie-Vergleiche modelliert. Diese Strukturen wurden anschließend für „rigid body docking“-Simulationen verwendet, die Interaktionspartner identifizierten und Information über die Gesamtstruktur des Komplexs lieferten.

References

6. References

Abascal, F. and Valencia, A. (2002) Clustering of proximal sequence space for the identification of protein families. Bioinformatics 18, 908-921.

Aloy, P. and Russell, R.B. (2004) Ten thousand interactions for the molecular biologist. Nat. Biotechnol. 22, 1317-1321.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. J. Mol. Biol. 215, 403-410.

Altschul, S.F, Madden, T.L, Schäffer, A.A, Zhang, J., Zhang, Z, Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.

Ames, B.D., Korman, T.P., Zhang, W., Smith, P., Vu, T., Tang, Y. and Tsai, S. (2008) Crystal structure and functional analysis of tetracenomycin ARO/CYC: implications for cyclization specificity of aromatic polyketides. Proc. Natl. Acad. Sci. U.S.A. 105, 5349-5354.

Andrusier, N., Mashiach, E., Nussinov, R. and Wolfson, H.J. (2008) Principles of flexible protein-protein docking. Proteins 73, 271-289.

Arnold, K., Bordoli, L., Kopp, J. and Schwede, T. (2006) The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. Bioinformatics 22, 195-201.

Arthur, C.J., Szafranska, A.E., Long, J., Mills, J., Cox, R.J., Findlow, S.C., Simpson, T.J., Crump, M.P. and Crosby, J. (2006) The malonyl transferase activity of type II polyketide synthase acyl carrier proteins. Chem. Biol. 13, 587-596.

Austin, M.B. and Noel, J.P. (2003) The chalcone synthase superfamily of type III polyketide synthases. Nat. Prod. Rep. 20, 79-110.

Austin, M.B., Saito, T., Bowman, M.E., Haydock, S., Kato, A., Moore, B.S., Kay, R.R. and Noel, J.P. (2006) Biosynthesis of *Dictyostelium discoideum* differentiation-inducing factor by a hybrid type I fatty acid-type III polyketide synthase. Nat. Chem. Biol. 2, 494-502.

Bachmann, B.O. (2005) Decoding chemical structures from genomes. Nat. Chem. Biol. 1, 244-245.

Baker, D. and Sali, A. (2001) Protein structure prediction and structural genomics. Science 294, 93-96.

Baker, D. (2006) Prediction and design of macromolecular structures and interactions. Phil. Trans. R. Soc. B 361, 459-463.

Bapteste, E., Brinkmann, H., Lee, J.A., Moore, D.V., Sensen, C.W., Gordon, P., Duruflé, L., Gaasterland, T., Lopez, P., Müller, M., *et al*. (2002) The analysis of 100 genes supports the grouping of three highly divergent amoebae: *Dictyostelium*, *Entamoeba*, and *Mastigamoeba*. Proc. Natl. Acad. Sci. U.S.A. 99, 1414-1419.

Bedford, D., Jacobsen, J.R., Luo, G., Cane, D.E. and Khosla, C. (1996) A functional chimeric modular polyketide synthase generated via domain replacement. Chem. Biol. 3, 827-831.

Bentley, S.D., Chater, K.F., Cerdeño-Tárraga, A., Challis, G.L., Thomson, N.R., James, K.D., Harris, D.E., Quail, M.A., Kieser, H., Harper, D., *et al*. (2002) Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). Nature 417, 141-147.

Besemer, J. and Borodovsky, M. (2005). GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. Nucleic Acids Res, 33 (Web Server issue), W451-4.

Birney, E., Clamp, M. and Durbin, R. (2004) GeneWise and Genomewise. Genome Res. 14, 988-995.

Bisang, C., Long, P.F., Cortés, J., Westcott, J., Crosby, J., Matharu, A.L., Cox, R.J., Simpson, T.J., Staunton, J. and Leadlay, P.F. (1999) A chain initiation factor common to both modular and aromatic polyketide synthases. Nature 401, 502-505.

Bowie, J.U., Lüthy, R. and Eisenberg, D. (1991) A method to identify protein sequences that fold into a known three-dimensional structure. Science 253, 164-170.

Brent, M.R. (2005) Genome annotation past, present, and future: how to define an ORF at each locus. Genome Res. 15, 1777-1786.

Brown, D.P., Krishnamurthy, N. and Sjölander, K. (2007) Automated protein subfamily identification and classification. PLoS Comput. Biol. 3, e160.

Burge, C. and Karlin, S. (1997) Prediction of complete gene structures in human genomic DNA. J. Mol. Biol. 268, 78-94.

Burson, K. and Khosla, C. (2000) Dissecting the Chain Length Specificity in Bacterial Aromatic Polyketide Synthases using Chimeric Genes. Tetrahedron 56, 9401-9408.

Cabin, R.J. and Mitchell, R.J. (2000) To Bonferroni or not to Bonferroni: when and how are the questions. ESA Bull. 81, 246-248.

Caffrey, P. (2003) Conserved amino acid residues correlating with ketoreductase stereospecificity in modular polyketide synthases. ChemBioChem 4, 654-657.

Cane, D.E., Walsh, C.T. and Khosla, C. (1998) Harnessing the biosynthetic bode: bombinations, permutations, and mutations. Science 282, 63-68.

Chakrabarti, S., Bryant, S.H. and Panchenko, A.R. (2007) Functional specificity lies within the properties and evolutionary changes of amino acids. J. Mol. Biol. 373, 801-810.

Chen, G., Zhuchenko, O. and Kuspa, A. (2007) Immune-like phagocyte activity in the social amoeba. Science 317, 678-681.

Chen, R., Li, L. and Weng, Z. (2003) ZDOCK: an initial-stage protein-docking algorithm. Proteins 52, 80-87.

Chan, Y.A., Podevels, A.M., Kevanya, B.M. and Thomas, M.G. (2009) Biosynthesis of polyketide synthase extender units. Nat. Prod. Rep. 26, 90-114.

Chothia, C. and Lesk, A.M. (1986) The relation between the divergence of sequence and structure in proteins. Embo J. 5, 823-826.

Comeau, S.R., Gatchell, D.W., Vajda, S. and Camacho, C.J. (2004) ClusPro: an automated docking and discrimination method for the prediction of protein complexes. Bioinformatics 20, 45-50.

Cortes, J., Wiesmann, K.E., Roberts, G.A., Brown, M.J., Staunton, J. and Leadlay, P.F. (1995) Repositioning of a domain in a modular polyketide synthase to promote specific chain cleavage. Science 268, 1487-1489.

Crump, M.P., Crosby, J., Dempsey, C.E., Parkinson, J.A., Murray, M., Hopwood, D.A. and Simpson, T.J. (1997) Solution structure of the actinorhodin polyketide synthase acyl carrier protein from *Streptomyces coelicolor* A3(2). Biochemistry 36, 6000-6008.

Curwen, V., Eyras, E., Andrews, T.D., Clarke, L., Mongin, E., Searle, S.M. and Clamp, M. (2004) The Ensembl automatic gene annotation system. Genome Res. 14, 942-950.

Demain, A.L. (2009) Antibiotics: natural products essential to human health. Med. Res. Rev. 29, 821-842.

Donald, J.E. and Shakhnovich, E.I. (2005) Determining functional specificity from protein sequences. Bioinformatics 21, 2629-2635.

Duhovny, D., Nussinov, R. and Wolfson, H. (2002) Efficient unbound docking of rigid molecules. In Algorithms in Bioinformatics. Springer-Verlag, London, p. 185–200.

Dunlap, W.C., Jaspars, M., Hranueli, D., Battershill, C.N., Perić-Concha, N., Zucko, J., Wright, S.H. and Long, P.F. (2006) New methods for medicinal chemistry - universal gene cloning and expression systems for production of marine bioactive metabolites. Curr. Med. Chem. 13, 697-710.

Eddy, S.R. (1996) Hidden Markov models. Curr. Opin. Struct. Biol. 6, 361-365.

Eddy, S.R. (1998) Profile hidden Markov models. Bioinformatics 14, 755-763.

Eddy, S.R. (2003) HMMER User's Guide - Biological sequence analysis using profile hidden Markov models. HHMI/Washington University School of Medicine.

Eichinger, L., Pachebat, J.A., Glöckner, G., Rajandream, M., Sucgang, R., Berriman, M., Song, J., Olsen, R., Szafranski, K., Xu,Q., *et al*. (2005) The genome of the social amoeba *Dictyostelium discoideum*. Nature 435, 43-57.

Eswar, N., Webb, B., Marti-Renom, M.A., Madhusudhan, M.S., Eramian, D., Shen, M., Pieper, U. and Sali, A. (2007) Comparative protein structure modeling using MODELLER. Curr. Protoc. Protein Sci. Chapter 2: Unit 2.9.

Feenstra, K.A., Pirovano, W., Krab, K. and Heringa, J. (2007) Sequence harmony: detecting functional specificity from alignments. Nucleic Acids Res. 35, W495-498.

Felsenstein, J. (2004) Inferring phylogenies 2nd. Sinauer Associates, Sunderland, MA.

Finn, R.D., Tate, J., Mistry, J., Coggill, P.C., Sammut, S.J., Hotz, H., Ceric, G., Forslund, K., Eddy, S.R., Sonnhammer, E.L., *et al*. (2008) The Pfam protein families database. Nucleic Acids Res. 36, D281-288.

Foerstner, K.U., Doerks, T., Creevey, C.J., Doerks, A. and Bork, P. (2008) A computational screen for type I polyketide synthases in metagenomics shotgun data. PloS One 3, e3515

Fraser, A.G. and Marcotte, E.M. (2004) A probabilistic view of gene function. Nature Genet. 36, 559-564.

Ghosh, R., Chhabra, A., Phatale, P.A., Samrat, S.K., Sharma, J., Gosain, A., Mohanty, D., Saran, S. and Gokhale, R.S. (2008) Dissecting the functional role of polyketide synthases in *Dictyostelium discoideum*: biosynthesis of the differentiation regulating factor 4-methyl-5-pentylbenzene-1,3-diol. J. Biol. Chem. 283, 11348-11354.

Ginalski, K. (2006) Comparative modeling for protein structure prediction. Curr. Opin. Struct. Biol. 16, 172-177.

Gokan, N., Kikuchi, H., Nakamura, K., Oshima, Y., Hosaka, K. and Kubohara, Y. (2005) Structural requirements of *Dictyostelium* differentiation-inducing factors for their stalk-cell-inducing activity in *Dictyostelium* cells and anti-proliferative activity in K562 human leukemic cells. Biochem. Pharmacol. 70, 676-685.

Grimm, A., Madduri, K., Ali, A. and Hutchinson, C.R. (1994) Characterization of the *Streptomyces peucetius* ATCC 29050 genes encoding doxorubicin polyketide synthase. Gene 151, 1-10.

Hadfield, A.T., Limpkin, C., Teartasin, W., Simpson, T.J., Crosby, J. and Crump, M.P. (2004) The crystal structure of the *act*III actinorhodin polyketide reductase: proposed mechanism for ACP and polyketide binding. Structure 12, 1865-1875.

Hannenhalli, S.S. and Russell, R.B. (2000) Analysis and prediction of functional sub-types from protein sequence alignments. J. Mol. Biol. 303, 61-76.

Haydock, S.F., Aparicio, J.F., Molnár, I., Schwecke, T., Khaw, L.E., König, A., Marsden, A.F., Galloway, I.S., Staunton, J. and Leadlay, P.F. (1995) Divergent sequence motifs correlated

with the substrate specificity of (methyl)malonyl-CoA:acyl carrier protein transacylase domains in modular polyketide synthases. FEBS Lett. 374, 246-248.

Herrmann, K. M. (1995). The Shikimate Pathway: Early Steps in the Biosynthesis of Aromatic Compounds. Plant cell 7, 907-919.

Hertweck, C., Luzhetskyy, A., Rebets, Y. and Bechthold, A. (2007) Type II polyketide synthases: gaining a deeper insight into enzymatic teamwork. Nat. Prod. Rep. 24, 162-190.

Hopwood, D.A. and Sherman, D.H. (1990) Molecular genetics of polyketides and its comparison to fatty acid biosynthesis. Annu. Rev. Genet. 24, 37-66.

Hranueli, D., Cullum, J., Basrak, B., Goldstein, P. and Long, P.F. (2005) Plasticity of the streptomyces genome-evolution and engineering of new antibiotics. Curr. Med. Chem. 12, 1697-1704.

Ikeda, H., Ishikawa, J., Hanamoto, A., Shinose, M., Kikuchi, H., Shiba, T., Sakaki, Y., Hattori, M. and Omura, S. (2003) Complete genome sequence and comparative analysis of the industrial microorganism *Streptomyces avermitilis*. Nat. Biotechnol. 21, 526-531.

Jacobsen, J.R., Hutchinson, C.R., Cane, D.E. and Khosla, C. (1997) Precursor-directed biosynthesis of erythromycin analogs by an engineered polyketide synthase. Science 277, 367-369.

Janin, J., Henrick, K., Moult, J., Eyck, L.T., Sternberg, M.J., Vajda, S., Vakser, I. and Wodak, S.J. (2003) CAPRI: a Critical Assessment of PRedicted Interactions. Proteins 52, 2-9.

Jenke-Kodama, H., Börner, T. and Dittmann, E. (2006) Natural biocombinatorics in the polyketide synthase genes of the actinobacterium *Streptomyces avermitilis*. PLoS Comput. Biol. 2, e132.

Jenke-Kodama, H., Sandmann, A., Müller, R. and Dittmann, E. (2005) Evolutionary implications of bacterial polyketide synthases. Mol. Biol. Evol. 22, 2027-2039.

Jones, D.T. (1999) GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. J. Mol. Biol. 287, 797-815.

Kalinina, O.V., Mironov, A.A., Gelfand, M.S. and Rakhmaninova, A.B. (2004) Automated selection of positions determining functional specificity of proteins by comparative analysis of orthologous groups in protein families. Protein Sci. 13, 443-456.

Kao, C.M., Luo, G., Katz, L., Cane, D.E. and Khosla, C. (1995) Manipulation of macrolide ring size by directed mutagenesis of a modular polyketide synthase. J. Am. Chem. Soc. 117, 9105-9106.

Karolchik, D., Kuhn, R.M., Baertsch, R., Barber, G.P., Clawson, H., Diekhans, M., Giardine, B., Harte, R.A., Hinrichs, A.S., Hsu, F., *et al*. (2008) The UCSC Genome Browser Database: 2008 update. Nucleic Acids Res. 36, D773-779.

Keating, T.A. and Walsh, C.T. (1999) Initiation, elongation, and termination strategies in polyketide and polypeptide antibiotic biosynthesis. Curr. Opin. Chem. Biol. 3, 598-606.

Keatinge-Clay, A. (2008) Crystal structure of the erythromycin polyketide synthase dehydratase. J. Mol. Biol. 384, 941-953.

Keatinge-Clay, A.T. and Stroud, R.M. (2006) The structure of a ketoreductase determines the organization of the beta-carbon processing enzymes of modular polyketide synthases. Structure 14, 737-748.

Keatinge-Clay, A.T., Maltby, D.A., Medzihradszky, K.F., Khosla, C. and Stroud, R.M. (2004) An antibiotic factory caught in action. Nat. Struct. Mol. Biol. 11, 888-893.

Keatinge-Clay, A.T., Shelat, A.A., Savage, D.F., Tsai, S.C., Miercke, L.J., O'Connell, J.D., Khosla, C. and Stroud, R.M. (2003) Catalysis, specificity, and ACP docking site of *Streptomyces coelicolor* malonyl-CoA:ACP transacylase. Structure 11, 147-154.

Kennedy, J., Auclair, K., Kendrew, S.G., Park, C., Vederas, J.C. and Hutchinson, C.R. (1999) Modulation of polyketide synthase activity by accessory proteins during lovastatin biosynthesis. Science 284, 1368-1372.

Khosla, C., Gokhale, R.S., Jacobsen, J.R. and Cane, D.E. (1999) Tolerance and specificity of polyketide synthases. Annu. Rev. Biochem. 68, 219-253.

Kroken, S., Glass, N. L., Taylor, J. W., Yoder, O. C. and Turgeon, B. G. (2003) Phylogenomic analysis of type I polyketide synthase genes in pathogenic and saprobic ascomycetes. Proc. Natl. Acad. Sci. U.S.A. 100, 15670-5.

Kwan, D.H., Sun, Y., Schulz, F., Hong, H., Popovic, B., Sim-Stark, J.C., Haydock, S.F. and Leadlay, P.F. (2008) Prediction and manipulation of the stereochemistry of enoylreduction in modular polyketide synthases. Chem. Biol. 15, 1231-1240.

Lau, J., Fu, H., Cane, D.E. and Khosla, C. (1999) Dissecting the role of acyltransferase domains of modular polyketide synthases in the choice and stereochemical fate of extender units. Biochemistry 38, 1643-1651.

Lautru, S., Deeth, R.J., Bailey, L.M. and Challis, G.L. (2005) Discovery of a new peptide natural product by *Streptomyces coelicolor* genome mining. Nat. Chem. Biol. 1, 265-269.

Li, Q., Khosla, C., Puglisi, J.D. and Liu, C.W. (2003) Solution structure and backbone dynamics of the holo form of the frenolicin acyl carrier protein. Biochemistry 42, 4648-4657.

Marti-Renom, M.A., Fiser, A., Madhusudhan, M.S., John, B., Stuart, A., Eswar, N., Pieper, U., Shen, M. and Sali, A. (2003) Modeling protein structure from its sequence. In Current Protocols in Bioinformatics. John Wiley & Sons, Inc., pp. 5.1.1-5.1.32.

Matharu, A.L., Cox, R.J., Crosby, J., Byrom, K.J. and Simpson, T.J. (1998) MCAT is not required for in vitro polyketide synthesis in a minimal actinorhodin polyketide synthase from *Streptomyces coelicolor*. Chem. Biol. 5, 699-711.

McDaniel, R., Thamchaipenet, A., Gustafsson, C., Fu, H., Betlach, M. and Ashley, G. (1999) Multiple genetic modifications of the erythromycin polyketide synthase to produce a library of novel "unnatural" natural products. Proc. Natl. Acad. Sci. U.S.A. 96, 1846-1851.

Menzella, H.G., Reid, R., Carney, J.R., Chandran, S.S., Reisinger, S.J., Patel, K.G., Hopwood, D.A. and Santi, D.V. (2005) Combinatorial polyketide biosynthesis by de novo design and rearrangement of modular polyketide synthase genes. Nat. Biotechnol. 23, 1171-1176.

Minowa, Y., Araki, M. and Kanehisa, M. (2007) Comprehensive analysis of distinctive polyketide and nonribosomal peptide structural motifs encoded in microbial genomes. J. Mol. Biol. 368, 1500-1517.

Moss, S.J., Martin, C.J. and Wilkinson, B. (2004) Loss of co-linearity by modular polyketide synthases: a mechanism for the evolution of chemical diversity. Nat. Prod. Rep. 21, 575-593.

Mount, D.W. (2004) Bioinformatics: Sequence and Genome Analysis 2 ed. Cold Spring Harbor Laboratory Press.

Newman, D.J. and Cragg, G.M. (2004) Marine natural products and related compounds in clinical and advanced preclinical trials. J. Nat. Prod. 67, 1216-1238.

Newman, D.J. and Cragg, G.M. (2007) Natural products as sources of new drugs over the last 25 years. J. Nat. Prod. 70, 461-477.

Oliynyk, M., Brown, M.J., Cortés, J., Staunton, J. and Leadlay, P.F. (1996) A hybrid modular polyketide synthase obtained by domain swapping. Chem. Biol. 3, 833-839.

Pazos, F., Rausell, A. and Valencia, A. (2006) Phylogeny-independent detection of functional residues. Bioinformatics 22, 1440-1448.

Pieper, U., Eswar, N., Webb, B.M., Eramian, D., Kelly, L., Barkan, D.T., Carter, H., Mankoo, P., Karchin, R., Marti-Renom, M.A., *et al*. (2009) MODBASE, a database of annotated comparative protein structure models and associated resources. Nucleic Acids Res. 37, D347-354.

Rausch, C., Weber, T., Kohlbacher, O., Wohlleben, W. and Huson, D. H. (2005) Specificity prediction of adenylation domains in nonribosomal peptide synthetases (NRPS) using transductive support vector machines (TSVMs). Nucleic Acids Res. 33, 5799-808.

Reeves, C.D., Murli, S., Ashley, G.W., Piagentini, M., Hutchinson, C.R. and McDaniel, R. (2001) Alteration of the substrate specificity of a modular polyketide synthase acyltransferase domain through site-specific mutations. Biochemistry 40, 15464-15470.

Reeves, G.A., Talavera, D. and Thornton, J.M. (2009) Genome and proteome annotation: organization, interpretation and integration. J. R. Soc. Interface 6, 129-147.

Reid, R., Piagentini, M., Rodriguez, E., Ashley, G., Viswanathan, N., Carney, J., Santi, D.V., Hutchinson, C.R. and McDaniel, R. (2003) A model of structure and catalysis for ketoreductase domains in modular polyketide synthases. Biochemistry 42, 72-79.

Ridley, C. P., Lee, H. Y. and Khosla, C. (2008. Evolution of polyketide synthases in bacteria. Proc. Natl. Acad. Sci. U.S.A. 105, 4595-4600.

Ritchie, D.W. (2008) Recent progress and future directions in protein-protein docking. Curr. Protein Pept. Sci. 9, 1-15.

Rivero, F. (2002) mRNA processing in *Dictyostelium*: sequence requirements for termination and splicing. Protist 153, 169-176.

Roberts, F., Roberts, C. W., Johnson, J. J., Kyle, D. E., Krell, T., Coggins, J. R., *et al*. (1998) Evidence for the shikimate pathway in apicomplexan parasites. Nature 393, 801-805.

Rusch, D.B., Halpern, A.L., Sutton, G., Heidelberg, K.B., Williamson, S., Yooseph, S., Wu, D., Eisen, J.A., Hoffman, J.M., Remington, K., *et al*. (2007) The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. PLoS. Biol. 5, e77.

Russell, R.B., Alber, F., Aloy, P., Davis, F.P., Korkin, D., Pichaud, M., Topf, M. and Sali, A. (2004) A structural perspective on protein-protein interactions. Curr. Opin. Struct. Biol. 14, 313-324.

Schneidman-Duhovny, D., Inbar, Y., Nussinov, R. and Wolfson, H.J. (2005) PatchDock and SymmDock: servers for rigid and symmetric docking. Nucleic Acids Res. 33, W363-367.

Serre, L., Verbree, E.C., Dauter, Z., Stuitje, A.R. and Derewenda, Z.S. (1995) The *Escherichia coli* malonyl-CoA:acyl carrier protein transacylase at 1.5-A resolution. Crystal structure of a fatty acid synthase component. J. Biol. Chem. 270, 12961-12964.

Shen, J., Xu, X., Cheng, F., Liu, H., Luo, X., Shen, J., Chen, K., Zhao, W., Shen, X. and Jiang, H. (2003) Virtual screening on natural products for discovering active compounds and target information. Curr. Med. Chem. 10, 2327-2342.

Shen, Y., Yoon, P., Yu, T.W., Floss, H.G., Hopwood, D. and Moore, B.S. (1999) Ectopic expression of the minimal *whiE* polyketide synthase generates a library of aromatic polyketides of diverse sizes and shapes. Proc. Natl. Acad. Sci. U.S.A. 96, 3622-3627.

Sherman, D.H. (2005) The Lego-ization of polyketide biosynthesis. Nat. Biotechnol. 23, 1083-1084.

Shu, Y.Z. (1998) Recent natural products based drug development: a pharmaceutical industry perspective. J. Nat. Prod. 61, 1053-1071.

Singh, S.B. and Pelaez, F. (2008) Biodiversity, chemical diversity and drug discovery. Prog. Drug Res. 65, 141, 143-174.

Smith, S. and Tsai, S. (2007) The type I fatty acid and polyketide synthases: a tale of two megasynthases. Nat. Prod. Rep. 24, 1041-1072.

Stajich, J.E., Block, D., Boulez, K., Brenner, S.E., Chervitz, S.A., Dagdigian, C., Fuellen, G., Gilbert, J.G., Korf, I., Lapp, H. *et al*. (2002) The Bioperl toolkit: Perl modules for the life sciences. Genome Res. 12, 1611-1618.

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

Starcevic, A., Jaspars, M., Cullum, J., Hranueli, D. and Long, P.F. (2007) Predicting the nature and timing of epimerisation on a modular polyketide synthase. ChemBioChem 8, 28-31.

Staunton, J. and Weissman, K.J. (2001) Polyketide biosynthesis: a millennium review. Nat. Prod. Rep. 18, 380-416.

Sultana, A., Kallio, P., Jansson, A., Wang, J., Niemi, J., Mäntsälä, P. and Schneider, G. (2004) Structure of the polyketide cyclase SnoaL reveals a novel mechanism for enzymatic aldol condensation. Embo J. 23, 1911-1921.

Tae, H., Kong, E. and Park, K. (2007) ASMPKS: an analysis system for modular polyketide synthases. BMC bioinformatics 8, 327.

Tae, H., Sohng, J. K. and Park, K. (2009) Development of an analysis program of type I polyketide synthase gene clusters using homology search and profile hidden Markov model. J Microbiol Biotechnol. 19, 140-146.

Tang, L., Yoon, Y.J., Choi, C.Y. and Hutchinson, C.R. (1998) Characterization of the enzymatic domains in the modular polyketide synthase involved in rifamycin B biosynthesis by *Amycolatopsis mediterranei*. Gene 216, 255-265.

Tang, Y., Tsai, S. and Khosla, C. (2003) Polyketide chain length control by chain length factor. J. Am. Chem. Soc. 125, 12708-12709.

Thompson, T.B., Katayama, K., Watanabe, K., Hutchinson, C.R. and Rayment, I. (2004) Structural and functional analysis of tetracenomycin F2 cyclase from *Streptomyces glaucescens*. A type II polyketide cyclase. J. Biol. Chem. 279, 37956-37963.

UniProt Consortium (2010) The Universal Protein Resource (UniProt) in 2010. Nucleic Acids Res. 38, D142-148.

Van Lanen, S.G. and Shen, B. (2006) Microbial genomics for the improvement of natural product discovery. Curr. Opin. Microbiol. 9, 252-260.

Wallace, I.M. and Higgins, D.G. (2007) Supervised multivariate analysis of sequence groups to identify specificity determining residues. BMC bioinformatics 8, 135.

Watson, J.D., Laskowski, R.A. and Thornton, J.M. (2005) Predicting protein function from sequence and structural data. Curr. Opin. Struct. Biol. 15, 275-284.

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

Wenzel, S.C. and Müller, R. (2005) Formation of novel secondary metabolites by bacterial multimodular assembly lines: deviations from textbook biosynthetic logic. Curr. Opin. Chem. Biol. 9, 447-458.

Wheeler, D.L., Barrett, T., Benson, D.A., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., DiCuccio, M., Edgar, R., Federhen, S., *et al*. (2007) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 35, D5-12.

Wilming, L.G., Gilbert, J.G., Howe, K., Trevanion, S., Hubbard, T. and Harrow, J.L. (2008) The vertebrate genome annotation (Vega) database. Nucleic Acids Res. 36, D753-760

Witkowski, A., Joshi, A.K. and Smith, S. (2004) Characterization of the beta-carbon processing reactions of the mammalian cytosolic fatty acid synthase: role of the central core. Biochemistry 43, 10458-10466.

Yadav, G., Gokhale, R.S. and Mohanty, D. (2003) Computational approach for prediction of domain organization and substrate specificity of modular polyketide synthases. J. Mol. Biol. 328, 335-363.

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

Zhang, Y. (2008) Progress and challenges in protein structure prediction. Curr. Opin. Struct. Biol. 18, 342-348.

Zucko, J., Starcevic, A., Diminic, J., Elbekali, M., Lisfi, M., Long, P.F., Cullum, J. and 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.

Acknowledgments

All scientific work included in this thesis was done in the Department of Genetics, Technische Universität Kaiserslautern and partly in the Section for bioinformatics, Faculty of Food Technology and Biotechnology, University of Zagreb under the guidance of Prof. Dr. John Cullum and Prof. Dr. Daslav Hranueli as a part of bilateral cooperation between Germany and Croatia on a scientific project: "*In silico* studies of recombination in modular biosynthetic clusters".

I would like to thank Prof. Dr. John Cullum and Prof. Dr. Daslav Hranueli for supervision and guidance during the last few years and for a lot of interesting topics discussed.

I'm grateful for my friends Antonio, Janko and Sim with whom I had pleasure to work and drink a lot of coffee and other substances.

Many thanks to Erna and Nina for making life in the lab and everyday routine exciting and unpredictable ☺. Thanks to Ingeborg for help in all sorts of administrative issues which I find difficult even in Croatian.

Thanks to Nora for big help in bringing this to the end, afternoon coffees (teas) which I really enjoyed and a lot of complaining in between ☺

Thanks to Sarah for long discussions, "lectures", walks and nice time spent together. Thanks to Dime, Elena and Vera for friendship, support and opportunity to use "local" languages far from home.

And finally thanks to my family for tolerating me for all these years… And now I'm back!

Appendices

All the appendices stated below are supplied on the CD rom.

1. The entire PhD thesis in PDF format.

2. The supplementary material from the paper: "Pavle Goldstein, Jurica Zucko, Dušica Vujaklija, Anita Krisko, Daslav Hranueli, Paul F. Long, Catherine Etchebest, Bojan Basrak and John Cullum. Clustering of protein domains for functional and evolutionary studies. *BMC Bioinformatics*, *10*, 335, 2009".

3. The supplementary material from the paper: "Starcevic, A., 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*, 6882-6892, 2008".

4. The supplementary material from the paper: "Gaetano Castaldo, Jurica Zucko, Sibylle Heidelberger, Dušica Vujaklija, Daslav Hranueli, John Cullum, Pakorn Wattana-Amorn, Matthew P. Crump, John Crosby and Paul F. Long. Proposed arrangement of proteins forming a bacterial type II polyketide synthase. *Chem. Biol.*, *15*, 1156-1165, 2008".

5. The supplementary material from the paper: "Jurica Zucko, Nives Skunca, Tomaz Curk, Blaz Zupan, Paul F. Long, John Cullum, Richard Kessin and Daslav Hranueli. Polyketide synthase genes and the natural products potential of *Dictyostelium discoideum*. *Bioinformatics*, *23*, 2543–2549, 2007".

CURRICULUM VITAE

PERSONAL INFORMATION

Name	**JURICA ŽUČKO**
Address	**KOŽINČEV BREG 39, 10090 ZAGREB, CROATIA**
Telephone	**+385 1 348 3216**
Mobile	**+385 98 221 007**
E-mail	**jzucko@gmail.com**
Nationality	Croatian
Date of birth	01.04.1979.

PROFESSIONAL EXPERIENCE

9/2005 – 11/2008	Novalis Ltd., Zagreb.
2/2008 – 12/2009	Scientific assistant in the Department of Genetics, University of Kaiserslautern.
12/2009 -	Junior Research Assistant at the Section for Bioinformatics, Faculty of Food Technology and Biotechnology, University of Zagreb.

EDUCATION

1993 – 1997	*Lucijan Vranjanin* High school
1998 – 2005	Faculty of Food Technology and Biotechnology, University of Zagreb Diploma thesis: Function predictions of polyketide synthase domains by Hidden Markov Models
2007 –	PhD studies at the Department of Genetics, University of Kaiserslautern

Hiermit versichere ich, die vorliegende Dissertation in der Abteilung Genetik der Universität Kaiserslautern selbständig durchgeführt und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet zu haben.

Kaiserslautern, Juli 2010

Jurica Žučko