Available online at www.sciencedirect.com

SciVerse ScienceDirect

Current Opinion in Biotechnology

Design-based re-engineering of biosynthetic gene clusters: plug-and-play in practice

Hans-Jörg Frasch[1], Marnix H Medema[1,2], Eriko Takano[1,4] and Rainer Breitling[2,3,4]

Synthetic biology is revolutionizing the way in which the biosphere is explored for natural products. Through computational genome mining, thousands of biosynthetic gene clusters are being identified in microbial genomes, which constitute a rich source of potential novel pharmaceuticals. New methods are currently being devised to prioritize these gene clusters in terms of their potential for yielding biochemical novelty. High-potential gene clusters from any biological source can then be activated by 'refactoring' their native regulatory machinery, replacing it by synthetic, orthogonal regulation and optimizing enzyme expression to function effectively in an industry-compatible target host. Various part libraries and assembly technologies have recently been developed which facilitate this process.

Addresses
[1] Department of Microbial Physiology, Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, Groningen, The Netherlands
[2] Groningen Bioinformatics Centre, Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, Groningen, The Netherlands
[3] Institute of Molecular, Cell and Systems Biology, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow, United Kingdom
[4] Faculty of Life Sciences, Manchester Institute of Biotechnology, University of Manchester, 131 Princess Street, Manchester M1 7ND, United Kingdom

Corresponding authors: Takano, Eriko (eriko.takano@manchester.ac.uk) and Breitling, Rainer (rainer.breitling@manchester.ac.uk)

Current Opinion in Biotechnology 2013, **24**:1144–1150

This review comes from a themed issue on **Pharmaceutical biotechnology**

Edited by **Ajikumar Parayil** and **Federico Gago**

For a complete overview see the Issue and the Editorial

Available online 27th March 2013

0958-1669/$ – see front matter,

http://dx.doi.org/10.1016/j.copbio.2013.03.006

Introduction

Today's healthcare would be unimaginable without the employment of natural products. Most of our drugs against bacterial or viral infections, cancers, parasites and other maladies are originally derived from the secondary metabolites of bacteria, fungi or plants. Ever since the introduction of next-generation sequencing, the amount of genome information for natural product-producing organisms has been increasing rapidly, leading to the identification of an unexpectedly large number of biosynthetic gene clusters: typically up to a few dozen per genome, tens of thousands in total. It has become clear that most of the products of these gene clusters are not produced at detectable levels under laboratory conditions [1–3]; the gene clusters are 'silent', 'sleeping' or 'cryptic'. Sequence-guided genome mining and heterologous expression of biosynthesis gene clusters have been successful in awakening some cryptic clusters through knocking out or overexpressing regulatory genes [4,5,6•], but no high-throughput methodology has been fully developed yet.

Arguably, the biggest hurdle to performing gene cluster characterization quickly and in large numbers has been the fact that gene clusters originate from many different organisms, most of which are difficult or impossible to culture or to manipulate genetically. Recently, we proposed a new strategy to overcome this challenge [7••]. According to this strategy, gene clusters are extracted from the sequence databases *en masse* and screened in high throughput, subjecting them to a standardized protocol of synthetic biology-guided re-engineering (refactoring) and heterologous expression in pre-optimized plug-and-play hosts [7••].

The concept of refactoring derives from software engineering [8]. It was first introduced in synthetic biology by Chan *et al.* [9], who partially redesigned phage T7 to make it more amenable to engineering. They reannotated the T7 genome, divided it into its fundamental 'parts', and then simplified the composition and ordering of the parts along the genome. Intriguingly, the process of refactoring can also be applied to entire gene clusters. This was pioneered by Temme *et al.* [10••], who refactored the entire 23.5-kb/20-gene nitrogen fixation cluster from *Klebsiella oxytoca*, redesigning the operon structures of the gene cluster and replacing all regulatory elements by synthetic regulatory parts that had been characterized and shown to work in the target host organism, orthogonally from its native regulation. The work of Watanabe *et al.* [11] has shown that the replacement of native promoters by synthetic ones can indeed be used to obtain successful heterologous expression of known compounds. If a standardized strategy could be developed to computationally identify those gene clusters from across the tree of life that have the highest potential to yield novel types of chemical structures and biological activities, the

refactoring process could be applied to allow high-throughput screening of these gene clusters in optimized hosts. Thus, microbial small molecules could yield a multitude of novel drug leads that could serve to combat bacterial infections, cancer and many other diseases. Here, we will outline the practical ramifications of the entire process of selecting and redesigning biosynthetic pathways, from the computationally aided selection of candidate pathways to the choice of a suitable host and the specific design choices in refactoring and heterologous expression.

Prioritizing pathways and gene clusters

A number of computational approaches have recently been developed for effectively mining the bulk of genomic data for biosynthetic gene clusters [12,13,14•]. Yet, the question remains how to prioritize the thousands of gene clusters that result from this exhaustive database search: even if refactoring can be standardized to achieve relatively high throughput, dozens — not thousands — of gene clusters can be characterized with it in the short term. Moreover, randomly trying out gene clusters would be a wasteful process if a lot of information is available to make a good pre-selection.

An effective strategy for prioritization, used in other genomic applications (e.g. [15]), is to aim for an unbiased sampling of diversity. Until now, the sampling of natural product diversity has instead been rather biased. Biochemically, certain chemical compound classes (such as macrolides) have been strongly over-represented in gene cluster characterization attempts. Phylogenetically, certain taxonomic groups (such as actinomycetes) have also been strongly overrepresented. Prioritization attempts could aim to reduce such biases, by sampling more neglected gene cluster families and organisms.

On the other hand, some genomes and certain gene cluster families will be more likely to encode compounds with clinically useful bioactivities such as antimicrobials: when trying to optimize the sampling of diversity, this diversity should therefore be measured multidimensionally (Figure 1). Dimensions of gene cluster diversity could include the sequence diversity of core domains (as used in the recent NaPDoS tool by Ziemert *et al.* [16•,17]), the overall architectural and sequence diversity of entire gene clusters (as measured by a distance metric, e.g. [18]), the combinatorial diversity of subclusters encoding specific chemical moieties [19], the taxonomic diversity of the organisms encoding the compound, and the diversity in ecological niches occupied by the source organisms. On the basis of such data, a mathematical algorithm could then be devised to identify a set of gene clusters that would most strongly complement the already characterized gene clusters. Such a set of high-priority gene clusters would constitute a good starting point for refactoring attempts to explore the uncharted territories of the chemical universe.

Of course, this diversity-enhancing sampling strategy can also be applied in a local fashion, for example, identifying maximally diverse sets of modifying genes or gene units associated with a specific biosynthetic class of clusters. In the short run, this locally targeted approach might be the more promising, as it increases the probability of hitting on relevant bioactivities by concentrating the search on variations of well-established chemical scaffolds with demonstrated clinical potential, while still expanding it far beyond what is accessible by traditional screening methods.

Finding suitable host organisms

Once a gene cluster has been selected, the first important decision is usually the choice of host organism for over-expression. Naturally, this host will be easy to manipulate genetically, offer a clean background by not producing too many similar compounds itself, and ideally will have a metabolic setup that allows high-volume diversion of fluxes toward the precursors of the expected end compound of the cluster of interest. The advantages and limitations of various well-established industrial hosts, including bacteria, such as *Streptomyces* spp., *Escherichia coli*, *Pseudomonas* spp., *Bacillus* spp., and fungi, such as *Penicillium* spp. and *Aspergillus* spp., for the heterologous production of polyketides, nonribosomal peptides and isoprenoids have recently been reviewed extensively [20]. The currently available host strains are generally good starting points, but are not necessarily optimal: during evolution, organisms are usually not optimized for maximal production titers of secondary metabolites. One strategy to obtain better hosts for a certain class of chemicals is to use organisms that originally produce similar compounds as a starting point: besides the metabolism and metabolite-dependent transcriptional regulation being already tuned to achieve at least moderate levels of compound, this also reduces concerns about self-resistance in the case of antimicrobial activity of the end product, since many antibiotic producers possess redundant resistance mechanisms [21]. Another strategy uses metabolic modeling to identify those organisms that have an overall metabolic network topology that is best suited for the production of certain compound classes [22]. These can then be engineered at the regulatory level to divert their metabolic fluxes in such a way as to actualize this — sometimes hidden — potential.

While engineering host strains derived from species that are not traditionally used in biotechnology may unlock much hidden potential in the long run, a large amount of additional engineering for desirable properties beyond secondary metabolite production will be necessary for such species, for example, to improve overall growth rate on cheap nutrient sources, to facilitate growth at high