Aromatic Hydrocarbon Metabolism by *Rhodococcus* sp. I24:

Computational, Biochemical and Transcriptional Analysis

By

Jefferson A. Parker

B.S. Biology
Florida Agricultural and Mechanical University, 1996

SUBMITTED TO THE DEPARTMENT OF BIOLOGY IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

FEBRUARY 2004

Signature of Author:_____
Department of Biology
January 8, 2004

Certified by:_____
Anthony J. Sinskey
Professor of Microbiology
Thesis Supervisor

Accepted by:_____
Alan D. Grossman
Professor of Biology
Co-Chairman, Committee for Graduate Students

Aromatic Hydrocarbon Metabolism by *Rhodococcus* sp. I24: Computational, Biochemical and Transcriptional Analysis

By

Jefferson A. Parker

Submitted to the Department of Biology
on January 8, 2004 in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy in Biology

ABSTRACT

Rhodococcus sp. I24 is a Gram-positive soil bacterium being developed for the manufacture of (-)cis-(1S,2R)-1-aminoindan-2-ol, a key precursor in the production of the HIV-1 protease inhibitor Crixivan™, from the aromatic hydrocarbon indene. *Rhodococcus* sp. I24 was grown by batch fermentation in the presence of naphthalene and indene to measure changes in gene expression and aromatic hydrocarbon metabolism with DNA microarray technology. Genes were selected for microarray analysis based on functional annotation assignments made by the Consensus Annotation by Phylogeny Anchored Sequence Alignment (CAPASA) program, a high throughput system for automated functional annotation assignment of DNA sequence similarity search results. CAPASA was validated by comparison to several methods of annotation, and the agreement to other methods ranged from 75-94%. Microarray results were analyzed by the newly described method of trigonometric deconvolution, a mathematical system for the measurement of changes in gene expression across multiple growth conditions with a minimal number of hybridizations. The combined analysis of aromatic metabolism and gene expression reveal the differential expression of multiple polycyclic aromatic hydrocarbon dioxygenases in a substrate and growth phase dependent manner.

Thesis Supervisor: Anthony J. Sinskey
Title: Professor of Microbiology

Table of Contents

CHAPTER I

Introduction

Biotechnology and Bioinformatics, A Historical Perspective

The biological sciences are in the throes of a paradigm shift in what it means to "study biology". The first tremors of this change occurred in 1953 with the description of the structure of DNA (Watson et al. 1953), and fully erupted with the publication of the draft of the human genome sequence (Lander et al. 2001; Venter et al. 2001). Many events over the last fifty years have all contributed to the advent of this new biology, where the focus of study is on an entire system instead of a single functional unit. The translation of the genetic code (Khorana 1959; Tener et al. 1959) allowed biologists to understand the relation between the information storage function of DNA and the biological activity of protein. DNA (Maxam et al. 1977; Sanger et al. 1977) and protein sequencing (Edman et al. 1967) technologies were invented, and eventually became inexpensive enough to become routine tools of analysis. The advent of these tools required the development of tools and methods for digital storage, tracking, and manipulation of large amounts of biological information. The solution to the data overflow came in the form of the first vacuum tube computer, invented by International Business Machines in 1952. The IBM 701 could perform 17,000 calculations per second (17 kHz) and was housed in a very large room by the government and large corporations, which were the only organizations able to afford it. In 2003, computers weighing less than eight pounds can perform more than 306,000,000 calculations per second (180,000 times faster than the IBM 701) at costs within the range of most university students. The advent of the low cost personal computer allowed early bioinformaticists to construct the foundation of what would become some of the most useful tools for recombinant DNA technology, rapid sequence

similarity searches like FASTA (Pearson 1990) and BLAST (Altschul et al. 1990), and the field of genomics itself.

The early co-evolution of DNA and protein sequencing technology and the computer was a serendipitous coincidence whose value can be measured by the history of sequence databases. The first protein sequence database was a book and sequence similarity searches were performed by hand (Smith 1990; Hagen 2000). These were soon followed by computerized versions, which eventually grew into the Genbank database of the National Center for Biotechnology Information (Dayhoff 1974; Smith 1990; Benson et al. 2003). The compilation of DNA and protein sequence data was rapidly followed by the development of sequence alignment algorithms by mathematicians (Needleman et al. 1970; Smith et al. 1981). The alignment of multiple protein sequences enabled the measurement of amino acid residue use and sequence specific change to understand the evolutionary relationship of proteins across phylogeny and function to construct the Point Accepted Mutation (PAM) and BLOcks Substitution Matrix (BLOSUM) residue substitution matrices used today to measure sequence divergence (Dayhoff 1974; Henikoff et al. 1992). Mathematics, computer science, and sequencing had given rise to the field of bioinformatics, while sequencing, recombinant DNA technology, and protein purification had fused into biotechnology. These fields continue to blur together today giving rise to analytical tools, biological reagents, and previously unimagined opportunities to manipulate biological systems for biocatalysis, environmental remediation, and therapeutics.

Biocatalysis, An Industrial Introduction

The biotechnology community has spent the last century developing and refining processes for the overproduction of almost every compound known in metabolism (Manual of Industrial Microbiology 1986; Burton et al. 2002). Many native products of microbial systems have long been the basis of billion dollar industries including amino acids for livestock feed (Guillouet et al. 1999), production of industrial grade ethanol, and manufacturing of antibiotics to treat disease (Glazer 1995).

Recombinant DNA technology opened the door to protein therapeutics for the novel treatment of human diseases, which is still in the early stages of realizing actual benefits almost 20 years later. As the science of biotechnology continues to influence the business of manufacturing, there are new tools and forces affecting the pursuit of efficiency. Biotechnological process development is poised to maximize production at a minimal cost. Through systematic analysis, every aspect of the system from raw material feed stock delivery and reagent selection to reactor system and product recovery is refined to meet these goals (Doran 1995; Thomas et al. 2002). The discipline of process development is a science unto itself, but it has long been the domain of engineers. The rise of biological systems as a tool for material manufacturing has introduced a need to cross the lines between engineering and biology.

The chemical manufacturing industry has long sought the commercial scale production of fine chemicals from low cost biological feedstock to eliminate the world's reliance on environmentally damaging and nonrenewable petrochemicals (Dua et al. 2002). In spite of the vast potential gains of biocatalyst technology, limited advances have been made in the actual application of biological systems to the development of

novel manufacturing solutions. Advances in biocatalytic manufacturing by companies such as BASF (Heidelberg, Germany; (Schmid et al. 2001) and Lonza (Basel, Switzerland; (Shaw et al. 2003) serve as examples to the industry that such biological tools can make for commercially viable processes. *Rhodococcus erythropolis* has been successfully utilized to produce chiral sulfoxides from crude oil, simultaneously reducing smog-producing pollutants and creating useful synthetic reagents (Shaw et al. 2003). Still, analysis of 134 industrial biotransformation reactions reveals that the majority of reactions are performed by hydrolases (44%) and redox systems (30%) (Shaw et al. 2003). The truth of the matter is that designing and engineering novel biological systems for the manufacture of xenobiotic compounds is not a trivial undertaking.

The primary biocatalysts being developed are purified enzymes engineered to have more desirable reaction properties. Development of such reagents involves design of *de novo* activities by rational protein design (Bolon et al. 2002), directed evolution of enzyme properties to increase activity, altering substrate utilization or engineering new functional activity by mutagenesis (Farinas et al. 2001; Zhao et al. 2002), site specific mutagenesis of active site residues (Panke et al. 2002), and engineering enzymes to function in nonaqueous solvents (Khmelnitsky et al. 1999). Multisubunit enzymes and oxidative reactions requiring cofactor recycling often require the use of whole cells grown in fermentation cultures (Thiry et al. 2002). Using whole cells has several advantages including the ability to grow cultures to high density before introducing the reaction substrate, easy recovery of product from liquid culture medium (if it is water soluble), and having an easily renewable source of catalyst (Schmid et al. 2001). Research is ongoing to improve overall production (Zhang et al. 2002) and solvent

tolerance (de Bont 1998) of whole cell systems. The main decision of whether to use purified enzymes or whole cells, how to configure the reactor system, and which purification processes to use all depend on the physical and chemical properties of the product, what biological systems are in place, and experience.

Chiral Synthons in Pharmaceutical Manufacturing

The stereospecific interaction of small molecule therapeutic agents is an area of intense interest in the pharmaceutical industry today. Racemic mixtures of an active compound can have decreased efficacy (Ariens 1993) or tragic side effects. These reasons, as well as growing regulatory pressures, have led to the increased production of single enantiomer drugs (Persidis 1997).

Crixivan™

Indinavir sulfate (Figure 1) is a member of the class of HIV protease inhibitors that prevent the intracellular cleavage of the viral polyprotein into active subunits (Vacca et al. 1994) required for the construction of an active virus particle. The five chiral centers of indinavir sulfate provide the potential for 32 stereoisomer configurations, only one of which is therapeutically active and sold as the drug Crixivan™ (Merck and Co., Whitehouse Station, NJ). Clinical studies have demonstrated that treatement with Crixivan™ in combination with two HIV-1 reverse transcriptase inhibitors reduces patient viral load in the blood to undetectable levels (Plosker et al. 1999).

Chemical Synthesis of (-)cis-(1S,2R)-1-aminoindan-2-ol

The challenge of developing a successful biocatalytic process is highlighted by the complexity of the chemical synthesis of (-)cis-(1S,2R)-1-aminoindan-2-ol (-)-CAI, a key component in the production of Crixivan™. Two of the five chiral centers of the final

product are contained in this intermediate, and a technically challenging synthesis reaction is employed to fulfill the manufacturing needs of scalability and downstream processing (Reider 1997). Stereospecific catalysis in a biphasic aqueous/ organic reactor system enables the efficient synthesis of the activated precursor 1,2-indene oxide (Senanayake et al. 1996; Hughes et al. 1997). The activated epoxide is aminated with acetonitrile through a Ritter-type reaction to from the final (-)-CAI product (Senanayake et al. 1995).

Salen-Mn(III) complexes were discovered in the early 1990's that could readily catalyze the epoxidation of alkenes by sodium hypochlorite in a stereospecific fashion with a regular enantiomeric excess (ee) on the order of 70% (Figure 2a). The stereospecificity of this reaction was increased by 20% for some substrates by modifying the side groups surrounding the catalytic metal center in a logical fashion to physically block undesirable substrate approach paths (Jacobsen et al. 1991).

The final product (salen)Mn(III)Cl (MnLCl, or the Jacobsen catalyst; Figure 2b) is able to catalyze the epoxidation of a range of olefins with product yields of 63-96% with 89-97% ee. Utilizing just 1.5 mol% MnLCl with 12% aqueous NaOCl in chlorobenzene, indene is converted to the 1,2-epoxide with a yield of 88% and 86% ee. In spite of the exceptional product yield, the Jacobsen salen catalyst had several shortcomings under these conditions including loss of 40% of catalyst per hour to degradation and only 70% utilization of the indene substrate after four hours of reaction (Senanayake et al. 1996). Catalyst instability and reaction completion issues were solved by the addition of 4-(3-phenylpropyl) pyridine N-oxide (P_3NO; Figure 2b). P_3NO improved catalyst stability by decreasing the degradation rate to about 5% per hour, thus allowing the reaction to run to

completion within two hours. P_3NO had the added effect of decreasing the amount of catalyst needed per reaction to only 0.25 mol%. The minimal improvement in product yield to 90% with 88% ee is overshadowed by the massive decrease in the amount of catalyst needed per reaction. These results were scalable to a multi-kilogram process (Senanayake et al. 1996). P_3NO has such dramatic effects in this reaction because it acts as a surfactant shuttle, carrying the active oxidant HOCl into the organic chlorobenzene layer to activate the catalyst (Figure 2c). Various kinetic studies revealed that this activation step is rate limiting in the reaction, independent of indene concentration (Hughes et al. 1997). P_3NO increases the active reactor volume to include the entire organic phase of the system; without it, the oxidation step would be limited the aqueous/ organic interface.

The amination of 1,2-indan oxide by a Ritter-type reaction with acetonitrile under acidic conditions has been shown to proceed by a mechanism that strictly maintains the stereochemistry at the C2 carbon (Senanayake et al. 1995; Senanayake et al. 1995). Strong acid is used to open the epoxide ring at C1 to form a reactive carbenium ion, which exists in equilibrium with a nitrilium intermediate with the acetonitrile. The reaction is driven by the formation of a stable cis-5,5-ring derived methyl oxazoline. Acid catalyzed hydrolysis releases the free cis-aminoindanol with a yield of 60-65% and 100% ee (Senanayake et al. 1995; Senanayake et al. 1995). Due to the lower yield of the Ritter reaction, the total yield of (-)-CAI from indene is on the order of 60% but the 88% ee of the epoxidation reaction is maintained. The complete (-) cis-(1S,2R)-1-aminoindan-2-ol is readily fed into the remaining synthesis of Crixivan^{TM} (Reider 1997). The combination of epoxidation and amination establishes two of the five chiral centers

in the Crixivan™ molecule; in addition it produces the toxic waste products chlorobenzene, P_3NO, and degradation products of the Jacobsen salen catalyst.

Biocatalytic Synthesis of (-)cis-(1S,2R)-1-aminoindan-2-ol

One possible biocatalyst for the production of (-)cis-(1S,2R)-1-aminoindan-2-ol is the polycyclic aromatic hydrocarbon (PAH) dioxygenase class of enzymes. These redox enzymes have been studied *in vivo* for decades (Butler et al. 1997) to analyze their ability to chemically activate the chemically stable aromatic hydrocarbon ring. The classical arrangement of these systems is a three-part electron transport system which shuttles electrons from NADH through the reductase and ferredoxin components, to a Rieske iron-sulfur center in the terminal oxygenase, which incorporates both atoms of dioxygen into the aromatic nucleus (Figure 3). Classically, this activation is followed by dehydrogenation and ring cleavage reactions eventually degrading the aromatic substrate to catechol; which is fed into the normal aromatic amino acid degradation metabolism. The substrate range and stereospecificity of the oxygenation products is determined by the configuration of the terminal oxygenase (Resnick 1996; Boyd et al. 1998). More importantly, unnatural substrates can be incompletely metabolized by these systems and used as a source of chiral synthons for incorporation in other synthetic reactions (O'Brien et al. 2002).

(i) *Pseudomonas putida*

Strains of *Pseudomonas putida* were known as early as 1993 that can metabolize toluene completely as a sole carbon source (Gibson 1993). Research scientists of the Merck Bioprocess Research and Development group employed mutation and selection methods to isolate *P. putda* strains that could metabolize indene to cis-(1S, 2R)-indandiol

in a toluene independent fashion in a two phase, aqueous/soybean oil fermentation reactor system (Connors et al. 1997). Cis-(1S, 2R)-indandiol can be chemically converted to (-)cis-(1S,2R)-1-aminoindan-2-ol by the same Ritter reaction with acetonitrile previously described. Further analysis of this system revealed the production of the monooxygenation products 1-indenol and 1-indanone, as well as the downstream dehydrogenase product 1-keto-2-hydroxyindan. At best, the *P. putida* system was able to produce cis-(1S, 2R)-indandiol at 220 mg/L with 95% ee. The low yield but high enantiomeric excess suggested that a cis-(1R, 2S)-indandiol specific dehydrogenase was responsible for resolving the racemic mixture to high stereopurity, but the loss of total product yield was unacceptably high.

(ii) *Eschericia coli*

Further advances in the biocatalytic production of cis-(1S, 2R)-indandiol focused on the development of recombinant *E. coli* expressing the *P. putida* toluene dioxygenase (TDO) genes (Reddy et al. 1999). Merck researchers were able to increase the total product yield of cis-(1S, 2R)-indandiol to 1200 mg/L with ee of 98% when scaled up to a 23L fermentation (again using the aqueous/ soybean oil two phase system) by eliminating the competing side reactions and dehydrogenase degradation. Regardless of the vast improvement in production yield, *E. coli* and *P. putida* have a fundamental flaw for use in indene bioconversion, both are sensitive to indene and its metabolites. However, Merck scientists were confident because "the well-developed genetic manipulation system for *E. coli* should greatly help to overcome these problems rapidly through approaches such as directed evolution and protein engineering" (Reddy et al. 1999). No

other reports on the development of indene metabolizing *E. coli* strains have been released from Merck.

(iii) *Rhodococcus* sp. I24 and B264-1

The greatest success of biocatalytic production of cis-(1S, 2R)-indandiol by Merck Bioprocess R&D was achieved with two strains of the genus *Rhodococcus* isolated by selection for growth on naphthalene or toluene as a sole carbon source (Chartrain et al. 1998). The strain *Rhodococcus* sp. B264-1 was found to contain only toluene degrading activity, but it was able to biotransform indene to cis-(1S, 2R)-indandiol with a total yield of 2.0 g/L with 99+ % ee in a 14 L fermentation system. The other strain, *Rhodococcus* sp. I24 was found to posses both naphthalene and toluene degrading activities and a cis-(1S, 2R)-indandiol production yield of 1.4 g/L trans-(1R, 2R)-indandiol with a greater than 98% ee, which can also serve as a precursor for (-)cis-(1S,2R)-1-aminoindan-2-ol (Buckland et al. 1999). Chartrain et al. (1998) were able to partially dissect the regulation systems employed by *Rhodococcus* sp. I24 for indene bioconversion through induction studies by growing the cells in the presence of naphthalene or toluene and adding indene to monitor metabolite production (Chartrain et al. 1998). They found that cells induced with naphthalene predominantly produced cis-(1R, 2S)-indandiol, while cells pre-induced with toluene produced cis-(1S, 2R)-indandiol. Additionally, the naphthalene induced cells immediately produced the trans-(1R, 2R)-indandiol product. As with the *P. putida*, *Rhodococcus* sp. I24 was found to also produce 1-indenol, 1-indanone, and 1-keto-2-hydroxyindan. The induction metabolite profile of this bacterium suggested the presence of a toluene inducible dioxygenase responsible for the cis-(1S,2R) indandiol product, a naphthalene inducible dioxygenase activity producing the cis-

(1R,2S) metabolite, and a naphthalene inducible monooxygenase responsible for the trans-(1R,2R) indandiol product by way of a spontaneously hydrolyzed epoxide intermediate (Chartrain et al. 1998). Later refinement of this model by scientists at the Massachusetts Institute of Technology (M.I.T.) clarified that both cis-indandiol products were further metabolized by stereospecific dehydrogenases to the final ketohydroxyindan (Treadway et al. 1999; Yanagimachi et al. 2001), while the trans-(1R,2R) indandiol was a metabolic end product as shown in Figure 4a.

Rhodococcus sp. I24 indene bioconversion at M.I.T. initially focused on genetic characterization (Treadway et al. 1999) and metabolic flux analysis (Yanagimachi et al. 2001) to identify targets for cloning and control of the indene metabolism network. Treadway et al. (1999) successfully employed a cosmid library screen to identify the naphthalene inducible dioxygenase (*nid*, Figure 4b) gene cluster responsible for synthesis of cis-(1R,2S) indandiol. Additionally, the *nid* gene cluster was also found to be responsible for the production of 1-indenol and 1-indanone.

Metabolic flux analysis is an analysis method to quantitate the flow of compounds through a metabolic network to determine the key points were perturbations can be applied to maximally alter the flow to desired products (Stephanopoulos 1998). Yanagimachi et al. (2001) used ^{14}C-radiolabeled indene to study the flow of indene metabolites through *Rhodococcus* sp. KY1, a spontaneous mutant of *Rhodococcus* sp. I24 which lost the toluene metabolism pathway. By measuring transmembrane uptake and secretion, as well as intracellular changes in metabolite concentration, Yanagimachi et al. (2001) were able to determine that the major route of indene metabolism in the KY1 strain was through the monooxygenation of indene to (1S,2R) indene oxide and a pH

dependent stereoselective spontaneous hydrolysis to cis-(1S,2R) and trans-(1R,2R) indandiol (Stafford et al. 2002).

***Rhodococcus* sp. I24 Genomic Analysis**

Pulsed Field Gel Electrophoresis

The spontaneous generation of the *Rhodococcus* sp. KY1 strain was a reproducible and stable phenomenon that resulted in the loss of naphthalene metabolism. Pulsed field gel electrophoresis (PFGE) analysis of *Rhodococcus* sp. I24 and KY1 revealed the presence of a 50 kb and 340 kb extrachromosomal element in the I24 strain, while the KY1 strain only possessed the 50 kb element (H. Priefert et al., manuscript in preparation). Southern blot analysis, promoter fusion studies, and transconjugation experiments with the smaller element revealed that the naphthalene metabolizing enzyme activities reside on the 50 kb element, while the toluene activities were carried on the 340 kb element (H. Priefert et al., manuscript in preparation). Sequence analysis and transconjugation studies strongly suggest that the extrachromosomal elements carry all genes necessary for the complete metabolism of naphthalene or toluene as a sole carbon source.

Genomic Sequencing

Integrated Genomics Inc. (IG; Chicago, IL) was engaged in 2000 to determine the entire genome sequence of *Rhodococcus* sp. I24 to better characterize the array of metabolic activities it contained. Genomic characterization was performed using a comparative analysis to determine the function of the putative open reading frames (Overbeek et al. 2003). The original annotation identified approximately 5500 open reading frames listed for the *Rhodococcus* sp. I24 genome, about half of which had no

functional annotation associated after being processed by automated sequence homology searches and metabolic reconstruction analysis (Overbeek et al. 2000; Osterman et al. 2003). In order to evaluate the methodologies utilized by Integrated Genomics, I performed a check of about 20 sequences by BLASTx sequence homology to personally evaluate what the functional assignments were. I disagreed with about half of the IG assignments based on the output of the BLASTx output. This led to a situation where previously characterized genes could still have no functional annotation assigned in cases where different analysis methods disagreed, and created a need to develop alternative methods of genome annotation. In order to confirm the function of the genome and supply functional information I developed a system to perform multiple BLASTx searches in parallel batches, and the results of these searches were evaluated *by hand* to assign function to all of the open reading frames of the genome.

Global Analysis of Indene Bioconversion by *Rhodococcus* sp. I24

Several advances occurred in parallel during the time following the manual annotation of the *Rhodococcus* sp. I24 genome. First, the ability to grow multiple fermentation cultures in parallel was introduced. The Infors Sixfors six vessel fermenter from Appropriate Technical Resources (Laurel, MD) allowed the relatively high throughput growth and analysis of multiple aromatic hydrocarbon metabolism fermentations simultaneously. Second, a batch BLAST system designed by our group for automated execution of sequence similarity searches was combined with a fully automated functional annotation system capable of determining a best functional description of a DNA sequence. Most importantly though, was the development and dissemination of DNA microarray technology.

Chemists, biologists, mathematicians, and computer scientists are redrawing the classical models of biotechnological research to create a new era of system wide analysis and functional genomics. The research presented in this thesis represents initial efforts in a global view of biocatalyst development (Figure 5) where genetic transcription is correlated with metabolic activity This model can be expanded to include whole genome microarrays and measurement of all metabolic activities of an organism in parallel and *in vivo*. The first question is "what genes are present in *Rhodococcus* sp. I24?". In chapter 2, I describe the development of an automated system for the large-scale functional annotation of DNA open reading frame sequences, the Consensus Annotation by Phylogeny Anchored Sequence Alignment (CAPASA) program. CAPASA analyzes the output of a translated DNA versus protein database alignment search to evaluate sequence homology, taxonomic similarity, and functional annotation relevance to determine the single best description of the query sequence. The second area of inquiry is focused on determining the metabolic profile of *Rhodococcus* sp. I24 activity on the aromatic hydrocarbons indene and naphthalene, and how the expression of a class of enzymes correlates to these. In chapter 3, targeted DNA microarrays of known and suspected aromatic hydrocarbon dioxygenases of *Rhodococcus* sp. I24 are used to analyze this transcription. Also, I introduce novel methodologies for the analysis of DNA microarray data to measure the actual changes in transcription between physiological conditions directly, as opposed to the current method of presenting the log2 fold induction. Together these methods reveal a two component regulatory mechanism of aromatic metabolism that is dependent on the growth phase of the culture, and the particular aromatic substrate present.

Biocatalysis for the production of novel small molecules that do not exist in nature requires the integration of multiple fields. Biology, chemistry, materials science, computation, informatics, and engineering will all play a role in developing the industrial systems of the future. The interdisciplinary nature of this work will force the fall of traditional barriers between previously distinct fields of study and lead to more customization of research training and cross-discipline collaboration to achieve the goals of individual researchers.

Figure 1. The structure of Crixivan^{TM} (Vacca et al. 1994).

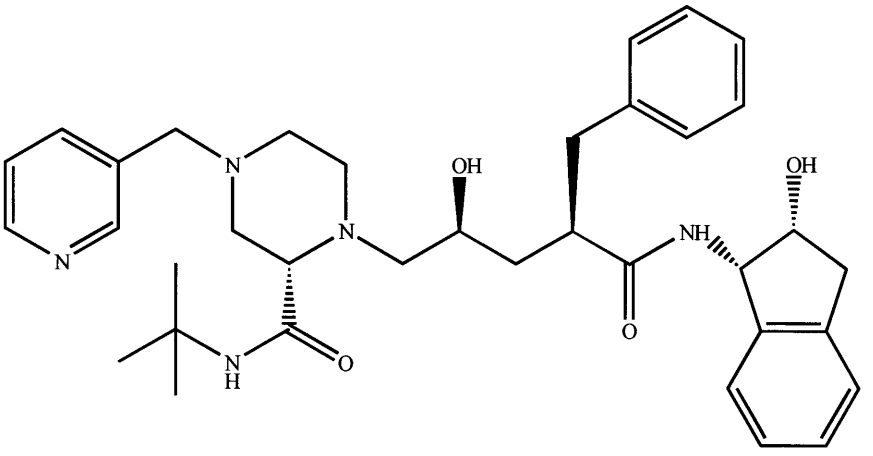


Figure 2a. The chemical reaction scheme to produce (-)cis-(1S,2R)-1-aminoindan-2-ol (far right) through the stereospecific epoxidation and amination of indene (left) through the reactive intermediate indene oxide (center) (Senanayake et al. 1995).

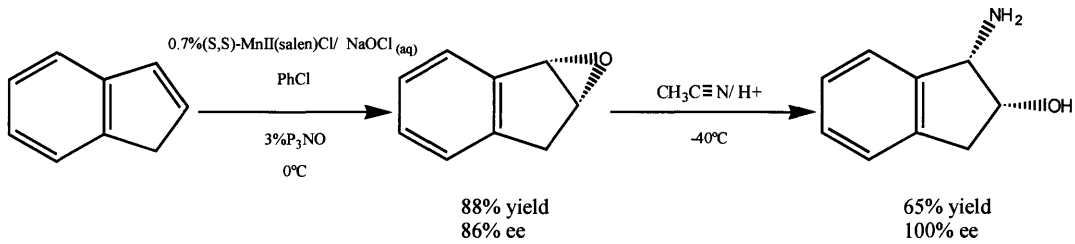


Figure 2b. The structure of the catalytic activation process of MnLCl by P_3NO (Senanayake et al. 1996). The large tertiary-butyl (t-Bu; $-C(CH_3)_3$) groups surrounding the reactive manganese core restrict the possible approach angles of substrate molecules, leading to the high stereospecificity of the catalyst species.

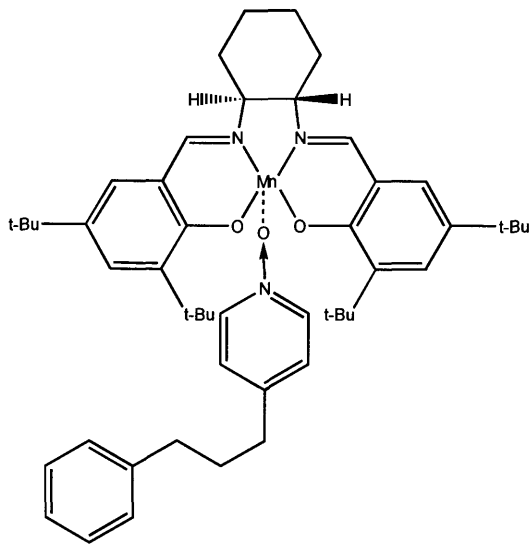


Figure 2c. The cyclic oxidation of indene by NaOCl through MnLCl in the biphasic reactor system resembles the enzymatic electron transfer system of bacterial oxygenase systems, shown in figure 3 (Senanayake et al. 1996).

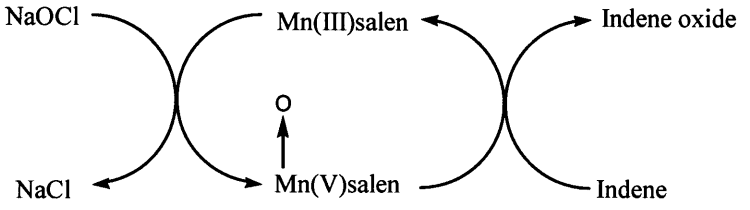


Figure 3. The enzymatic oxidation of polycyclic aromatic hydrocarbons is achieved by transporting electrons from $NADH+H^+$ through a three-subunit enzyme complex to the terminal dioxygense. The dioxygense subunit contains a Rieske Fe-S center, which serves as the electron acceptor and activation center for molecular oxygen. The activated oxygen molecule attacks the aromatic nucleus of the substrate through an enzyme-coordinated mechanism.

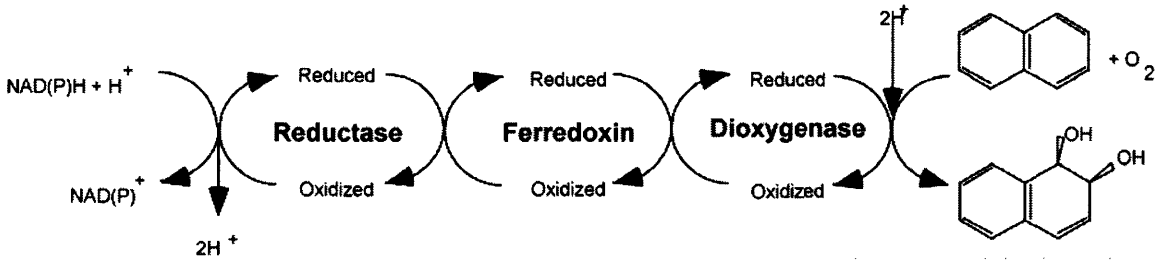


Figure 4a. *Rhodococcus* sp. I24 is able to metabolize indene to multiple products including cis-(IS, 2R)-indandiol and trans-(1R, 2R)-indandiol, both of which can serve as precursors for cis-(1S)-amino-(2R)-indanol (modified from Treadway et al. 1999).

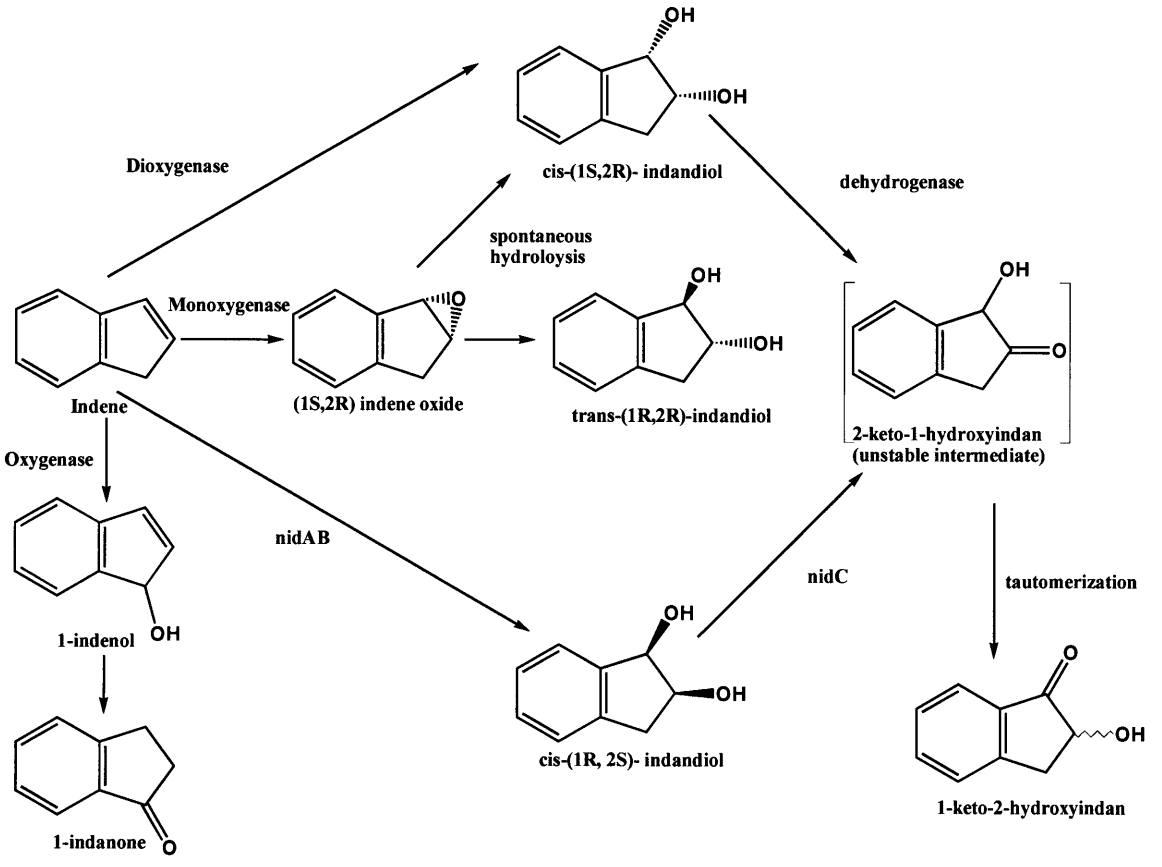


Figure 4b. The chromosome structure of several dioxygenase systems from Rhodococcussp. I24. Groups contain a large (*nidA, tidA*) and small (*nidB, tidB, nimA*) terminal dioxygenase subunit, and a dehydrogenase (*nidC, nimC*). The naphthalene inducible monooxygenase is characterized by its single subunit putative monooxygenase enzyme (*nimB*) (Treadway et al. 1999).

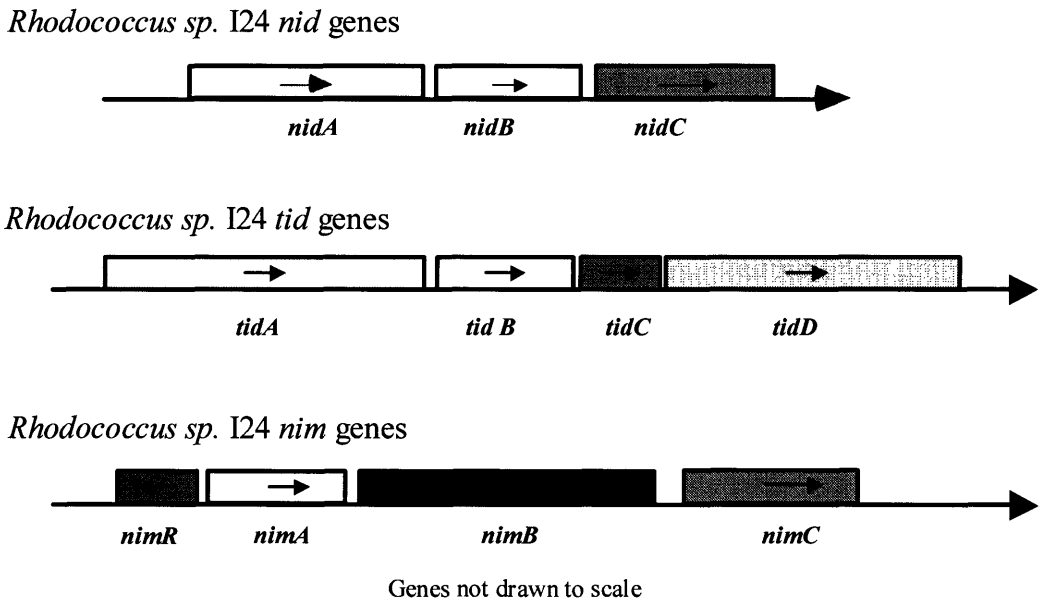
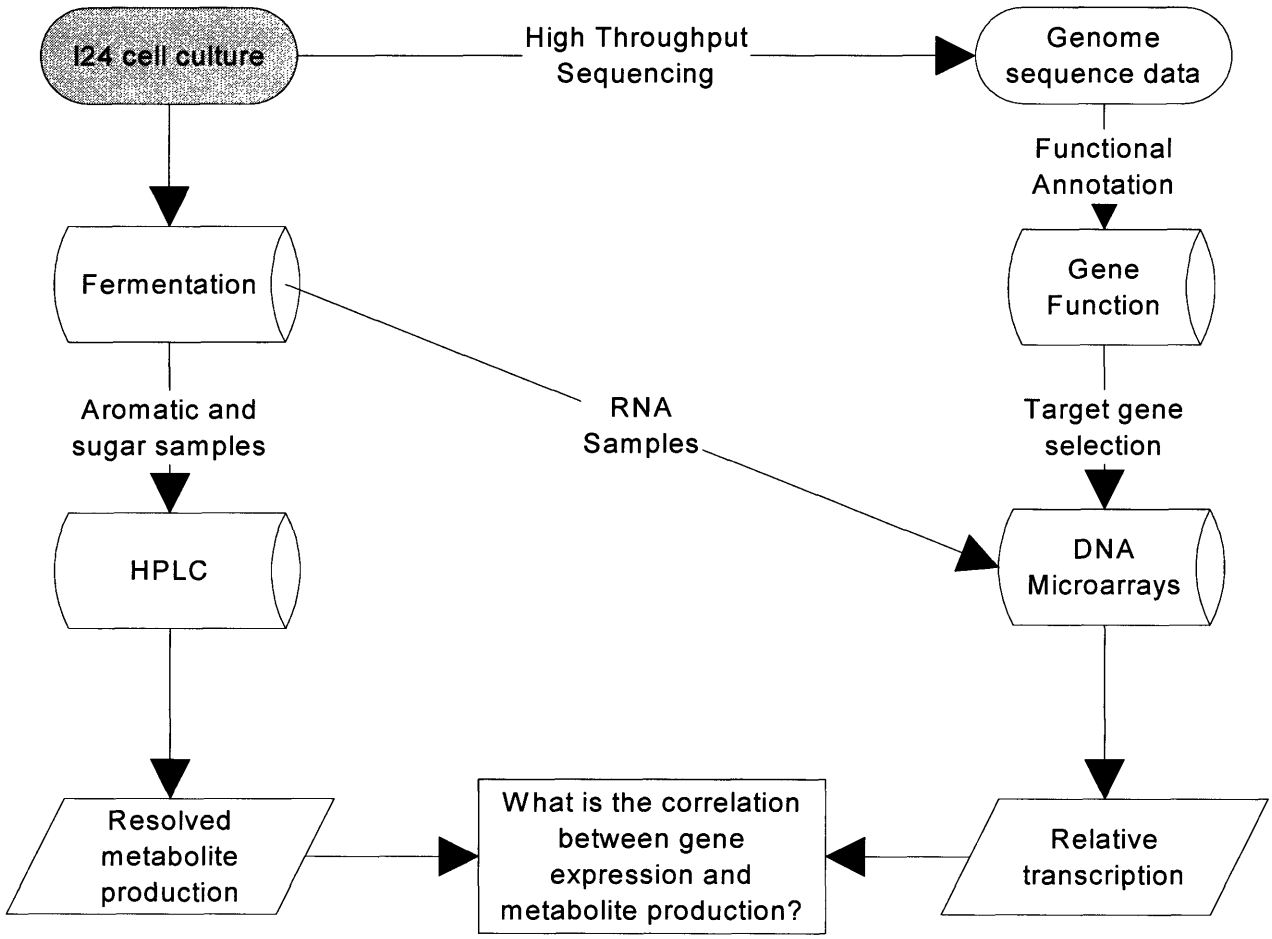


Figure 5. The global analysis plan for analysis of aromatic hydrocarbon metabolism in *Rhodococcus* sp. I24 from genome sequence and aromatic hydrocarbon metabolism to DNA microarrays and measurement of gene expression changes across multiple physiological conditions.



References

(1986). Manual of Industrial Microbiology and Biotechnology. Washington, DC, American Society for Microbiology.

Altschul, SF, W Gish, W Miller, EW Myers and DJ Lipman (1990). Basic local alignment search tool. J Mol Biol **215**(3): 403-10.

Ariens, EJ (1993). Nonchiral, Homochiral and Composite Chiral Drugs. Trends in Pharmacological Sciences **14**(2): 68-75.

Benson, DA, I Karsch-Mizrachi, DJ Lipman, J Ostell and DL Wheeler (2003). GenBank. Nucleic Acids Res **31**(1): 23-7.

Bolon, DN, CA Voigt and SL Mayo (2002). De novo design of biocatalysts. Curr Opin Chem Biol **6**(2): 125-9.

Boyd, DR and GN Sheldrake (1998). The dioxygenase-catalysed formation of vicinal cis-diols. Natural Product Reports **15**(3): 309-324.

Buckland, BC, SW Drew, NC Connors, MM Chartrain, C Lee, PM Salmon, K Gbewonyo, W Zhou, P Gailliot, R Singhvi, et al. (1999). Microbial conversion of indene to indandiol: a key intermediate in the synthesis of CRIXIVAN. Metab Eng **1**(1): 63-74.

Burton, SG, DA Cowan and JM Woodley (2002). The search for the ideal biocatalyst. Nat Biotechnol **20**(1): 37-45.

Butler, CS and JR Mason (1997). Structure-function analysis of the bacterial aromatic ring-hydroxylating dioxygenases. Adv Microb Physiol **38**: 47-84.

Chartrain, M, B Jackey, C Taylor, V Sandford, K Gbewonyo, L Lister, L Dimichele, C Hirsch, B Heimbuch, C Maxwell, D Pascoe, B Buckland, R Greasham (1998). Bioconversion of indene to *cis* (1S,2R) indandiol and *trans* (1R,2R) indandiol by *Rhodococcus* species. Journal of Fermentation and Bioengineering **86**(6): 550-558.

Connors, N, R Prevoznak, M Chartrain, J Reddy, R Singhvi, Z Patel, R Olewinski, P Salmon, J Wilson and R Greasham (1997). Conversion of indene to cis-(1S),(2R)-indandiol by mutants of *Pseudomonas putida* F1. Journal of Industrial Microbiology & Biotechnology **18**(6): 353-359.

Dayhoff, MO (1974). Computer analysis of protein sequences. Fed Proc **33**(12): 2314-6.

de Bont, JAM (1998). Solvent-tolerant bacteria in biocatalysis. Trends in Biotechnology **16**(12): 493-499.

Doran, PM (1995). Bioprocess Engineering Principles. London, UK, Academic Press.

Dua, M, A Singh, N Sethunathan and AK Johri (2002). Biotechnology and bioremediation: successes and limitations. Appl Microbiol Biotechnol **59**(2-3): 143-52.

Edman, P and G Begg (1967). A protein sequenator. Eur J Biochem **1**(1): 80-91.

Farinas, ET, T Bulter and FH Arnold (2001). Directed enzyme evolution. Curr Opin Biotechnol **12**(6): 545-51.

Gibson, DT (1993). Biodegradation, Biotransformation and the Belmont. Journal of Industrial Microbiology **12**(1): 1-12.

Glazer AN, H Nikaido (1995). Microbial Biotechnology: Fundamentals of Applied Microbiology. New York, NY, Freeman.

Guillouet, S, AA Rodal, G An, PA Lessard and AJ Sinskey (1999). Expression of the *Escherichia coli* catabolic threonine dehydratase in *Corynebacterium glutamicum* and its effect on isoleucine production. Applied and Environmental Microbiology **65**(7): 3100-7.

Hagen, JB (2000). The origins of bioinformatics. Nat Rev Genet **1**(3): 231-6.

Henikoff, S and JG Henikoff (1992). Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A **89**(22): 10915-9.

Hughes, DL, GB Smith, J Liu, GC Dezeny, CH Senanayake, RD Larsen, TR Verhoeven and PJ Reider (1997). Mechanistic Study of the Jacobsen Asymmetric Epoxidation of Indene. J Org Chem **62**(7): 2222-2229.

Jacobsen, EN, W Zhang, AR Muci, JR Ecker and L Deng (1991). Highly Enantioselective Epoxidation Catalysts Derived from 1,2-Diaminocyclohexane. Journal of the American Chemical Society **113**(18): 7063-7064.

Khmelnitsky, YL and JO Rich (1999). Biocatalysis in nonaqueous solvents. Curr Opin Chem Biol **3**(1): 47-53.

Khorana, HG (1959). Synthesis and structural analysis of polynucleotides. J Cell Comp Physiol **54**: 5-15.

Lander, ES, LM Linton, B Birren, C Nusbaum, MC Zody, J Baldwin, K Devon, K Dewar, M Doyle, W FitzHugh, et al. (2001). Initial sequencing and analysis of the human genome. Nature **409**(6822): 860-921.

Maxam, AM and W Gilbert (1977). A new method for sequencing DNA. Proc Natl Acad Sci U S A **74**(2): 560-4.

Needleman, SB and CD Wunsch (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol **48**(3): 443-53.

O'Brien, XM, JA Parker, PA Lessard and AJ Sinskey (2002). Engineering an indene bioconversion process for the production of cis-aminoindanol: a model system for the production of chiral synthons. Appl Microbiol Biotechnol **59**(4-5): 389-99.

Osterman, A and R Overbeek (2003). Missing genes in metabolic pathways: a comparative genomics approach. Curr Opin Chem Biol **7**(2): 238-51.

Overbeek, R, N Larsen, GD Pusch, M D'Souza, E Selkov, Jr., N Kyrpides, M Fonstein, N Maltsev and E Selkov (2000). WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. Nucleic Acids Res **28**(1): 123-5.

Overbeek, R, N Larsen, T Walunas, M D'Souza, G Pusch, E Selkov, Jr., K Liolios, V Joukov, D Kaznadzey, I Anderson, et al. (2003). The ERGO genome analysis and discovery system. Nucleic Acids Res **31**(1): 164-71.

Panke, S and MG Wubbolts (2002). Enzyme technology and bioprocess engineering. Curr Opin Biotechnol **13**(2): 111-6.

Pearson, WR (1990). Rapid and Sensitive Sequence Comparison with Fastp and Fasta. Methods in Enzymology **183**: 63-98.

Persidis, A (1997). Chiral-based therapeutics. Nat Biotechnol **15**(6): 594-5.

Plosker, GL and S Noble (1999). Indinavir: a review of its use in the management of HIV infection. Drugs **58**(6): 1165-203.

Reddy, J, C Lee, M Neeper, R Greasham and J Zhang (1999). Development of a bioconversion process for production of cis-1S,2R- indandiol from indene by

recombinant *Escherichia coli* constructs. Appl Microbiol Biotechnol **51**(5): 614-20.
Reider, PJ (1997). Advances in AIDS chemotherapy: The asymmetric synthesis of CRIXIVAN(R). Chimia **51**(6): 306-308.
Resnick, SM, David T. Gibson (1996). Oxidation of 6,7-Dihydro-5H-Benzocycloheptene by Bacterial Strains Expressing Napthalene Dioxygenase, Biphenyl Dioxygenase, and Toluene Dioxygenase Yields Homochiral Monol or cis-Diol Enantiomers as Major Products. Applied and Environmental Microbiology **62**(4): 1364-1368.
Sanger, F, S Nicklen and AR Coulson (1977). DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A **74**(12): 5463-7.
Schmid, A, JS Dordick, B Hauer, A Kiener, M Wubbolts and B Witholt (2001). Industrial biocatalysis today and tomorrow. Nature **409**(6817): 258-68.
Senanayake, CH, LM Dimichele, J Liu, LE Fredenburgh, KM Ryan, FE Roberts, RD Larsen, TR Verhoeven and PJ Reider (1995). Regiocontrolled and Stereocontrolled Syntheses of Cyclic Chiral Cis-Amino Alcohols from 1,2-Diols or Epoxides. Tetrahedron Letters **36**(42): 7615-7618.
Senanayake, CH, FE Roberts, LM Dimichele, KM Ryan, J Liu, LE Fredenburgh, BS Foster, AW Douglas, RD Larsen, TR Verhoeven, et al. (1995). The Behavior of Indene Oxide in the Ritter Reaction - a Simple Route to Cis-Aminoindanol. Tetrahedron Letters **36**(23): 3993-3996.
Senanayake, CH, GB Smith, KM Ryan, LE Fredenburgh, J Liu, FE Roberts, DL Hughes, RD Larsen, TR Verhoeven and PJ Reider (1996). The role of 4-(3-phenylpropyl)pyridine N-oxide (P3NO) in the manganese-salen-catalyzed asymmetric epoxidation of indene. Tetrahedron Letters **37**(19): 3271-3274.
Shaw, NM, KT Robins and A Kiener (2003). Lonza: 20 years of biotransformations. Advanced Synthesis & Catalysis **345**(4): 425-435.
Smith, TF (1990). The history of the genetic sequence databases. Genomics **6**(4): 701-7.
Smith, TF and MS Waterman (1981). Identification of common molecular subsequences. J Mol Biol **147**(1): 195-7.
Stafford, DE, KS Yanagimachi, PA Lessard, SK Rijhwani, AJ Sinskey and G Stephanopoulos (2002). Optimizing bioconversion pathways through systems analysis and metabolic engineering. Proc Natl Acad Sci U S A **99**(4): 1801-6.
Stephanopoulos GN, AA Aristidou, J Nielsen (1998). Metabolic Engineering: Principles and Methodologies. San Diego, CA, Academic Press.
Tener, GM, PT Gilham, WE Razzell, AF Turner and HG Khorana (1959). Studies on the chemical synthesis and enzymatic degradation of desoxyribo-oligonucleotides. Ann N Y Acad Sci **81**: 757-75.
Thiry, M and D Cingolani (2002). Optimizing scale-up fermentation processes. Trends in Biotechnology **20**(3): 103-105.
Thomas, SM, R DiCosimo and V Nagarajan (2002). Biocatalysis: applications and potentials for the chemical industry. Trends Biotechnol **20**(6): 238-42.
Treadway, SL, KS Yanagimachi, E Lankenau, PA Lessard, G Stephanopoulos and AJ Sinskey (1999). Isolation and characterization of indene bioconversion genes from *Rhodococcus* strain I24. Appl Microbiol Biotechnol **51**(6): 786-93.
Vacca, JP, BD Dorsey, WA Schleif, RB Levin, SL McDaniel, PL Darke, J Zugay, JC Quintero, OM Blahy, E Roth, et al. (1994). L-735,524: an orally bioavailable

human immunodeficiency virus type 1 protease inhibitor. Proc Natl Acad Sci U S A **91**(9): 4096-100.

Venter, JC, MD Adams, EW Myers, PW Li, RJ Mural, GG Sutton, HO Smith, M Yandell, CA Evans, RA Holt, et al. (2001). The sequence of the human genome. Science **291**(5507): 1304-51.

Watson, JD and FH Crick (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature **171**(4356): 737-8.

Yanagimachi, KS, DE Stafford, AF Dexter, AJ Sinskey, S Drew and G Stephanopoulos (2001). Application of radiolabeled tracers to biocatalytic flux analysis. Eur J Biochem **268**(18): 4950-60.

Zhang, YX, K Perry, VA Vinci, K Powell, WP Stemmer and SB del Cardayre (2002). Genome shuffling leads to rapid phenotypic improvement in bacteria. Nature **415**(6872): 644-6.

Zhao, H, K Chockalingam and Z Chen (2002). Directed evolution of enzymes and pathways for industrial biocatalysis. Curr Opin Biotechnol **13**(2): 104-10.

CHAPTER II

Consensus Annotation by Phylogeny Anchored Sequence Alignment (CAPASA)

Abstract

Recent advances in DNA sequencing technology, targeted DNA microarray design, and metagenomic analysis of uncultured microorganisms have revealed the need for automated methods of functional annotation. We have developed the Consensus Annotation by Phylogeny Anchored Sequence Alignment (CAPASA) for the analysis of sequence similarity search results and automated annotation of the query sequence. Search results are parsed into filtered training sets, quantified by sequence similarity scoring metrics and the collective agreement of taxonomic relationships and functional nomenclature between multiple search results. CAPASA was validated by comparison to several methods of annotation, and the agreement to other methods ranged from 75-94%. CAPASA is a system for the rapid functional annotation of DNA or protein sequences amenable to the average computer user and suitable for whole genome analysis.

The resources and methods described in this manuscript utilize public databases made available through the National Center for Biotechnology Information, the National Library of Medicine, and National Institutes of Health via the Genbank sequence databases, BLAST sequence alignment search programs, and taxonomy database. The source code of the CAPASA program and supplementary tables of results, are available for download at http://web.mit.edu/biology/sinskey/www/home.html in the publications section.

Introduction

Functional annotation refers to the characterization and assignment of biological activity to a gene product by experimentation, protein domain activity prediction, sequence homology, or many other methods for inferring the activity of a particular

biomolecule. Experimental validation is the most reliable of these methods, but this is a nontrivial undertaking with extreme cost both in terms of monetary consideration and in the time and labor associated. Automated sequence database searching that transfers the function of the highest similarity match to the query is a common practice, which is fast and simple to execute but often results in dubious functional assignments with no measure of annotation confidence (Brenner 1999; Koski et al. 2001). Choosing a method of functional annotation often involves a balanced choice by the researcher based on cost, time, computational resources, and familiarity with the different systems.

Advances in DNA sequencing technology and computational analysis have enabled the completion of the first draft of the full human genome, as well as the genome of rice, fruit fly, *Escherichia coli*, yeast, and many others (Goffeau et al. 1996; Blattner et al. 1997; Adams et al. 2000; Lander et al. 2001; Venter et al. 2001; Yu et al. 2002). The number of publicly available DNA sequences has grown at an exponential rate over the past decade. The high cost of DNA microarray technology has made the design and analysis of targeted gene sets involved in the regulatory and metabolic pathways of interest an attractive alternative to full genome analysis. Metagenomic analysis of uncultured environmental microorganisms relies solely on DNA sequence extracted from environmental samples (Schloss et al. 2003). Automated, scalable, and reliable methods are needed to assign putative functions to newly determined sequences. Such methods, if accessible to the research community at large with minimal cost in equipment, expertise, and time, would increase the speed at which novel genomes can be meaningfully interrogated.

We have developed an automated system for assigning functional annotations, implemented in the PERL scripting language: Consensus Annotation by Phylogeny Anchored Sequence Alignment (CAPASA). The flexible modular system architecture of our software streamlines the execution and analysis of a sequence alignment search output to select the best functional assignment to a query sequence from the search results relying on several parameters meant to mimic discriminations made in manual sequence similarity based annotation. CAPASA quantifies the sequence similarity, organism phylogenetic relatedness (using taxonomic lineage as a rapid approximation), and the name components of functional assignments within the output of a BLAST sequence similarity search (Altschul et al. 1990; Benson et al. 2003).

Methods

CAPASA release 1.0 (development version 8.0) was scripted using ActivePerl version 5.8.0.805 in the Windows 2000 OS environment. Full source code and installation instructions are available for download at http://web.mit.edu/biology/sinskey/www/home.html in the publications section. All processing was run on a DELL (Round Rock, TX) Dimension 4100 model XPS-Z with 933 MHz Pentium IIIc processor and 512 Mb DIMM RAM during evening hours (EST) to reduce the burden on NCBI computational resources.

BLASTx Sequence Alignment Search

The BLASTx search for CAPASA queued to the National Center for Biotechnology Information (NCBI) BLAST server from a local host via the QBlast URL API with gap introduction cost of 11, gap extension cost of 1, BLOSUM 62 substitution matrix, a maximum E-score of $1x10^{-20}$, and low complexity sequence filtering with the output in

HTML format and no graphical table (http://www.ncbi.nlm.nih.gov/BLAST/Doc/urlapi.html).

Results

CAPASA Software development

CAPASA was developed for the Windows operating system using the PERL scripting language (Wall 2000). The CAPASA algorithm quantifies the output of a BLASTx translated DNA versus protein database sequence similarity search (Gish et al. 1993). A feeder module of our software allows automated processing batched sequences in FASTA format, or HTML formatted BLAST search result files. The parameters of the BLASTx search are designed to reduce the selection of low homology and unrelated sequences from the Genbank non-redundant (nr) protein sequence database. Each alignment (x) will be associated with zero to several entries (y), each from a different contributing database. Each BLASTx alignment has both a bit score (S′) which is a measure of absolute sequence similarity, and E-score (E) an approximation of the probability of finding a better scoring sequence where:

$$S' = \frac{\lambda S - \ln K}{\ln 2} \qquad \textbf{Eq. 1}$$

and

$$E = mn2^{-S'} \qquad \textbf{Eq. 2}$$

The actual probability of such an alignment occurring is:

$$P = 1 - e^{-E} \qquad \textbf{Eq. 3}$$

(See Altschul et al. 1990; Karlin et al. 1990; Altschul 1991 and Gish et al. 1993 for a thorough discussion of BLAST scoring parameters). Sequence alignments with lower E-

scores have a high probability that there is not another sequence in the database that would match as well.

The CAPASA annotation score (α) of an entry is determined from three BLAST derived factors: the homology score (η) of the sequence, the taxonomy score (τ) of the organism which is an approximation of phylogenetic relatedness, and the consensus score (γ) of the function name of the entry (Figure 1)

$$\alpha_{xy} = \eta_x + \tau_y + \gamma_y \qquad \textbf{Eq. 4}$$

The scoring of each component is elaborated in the following sections.

CAPASA parses the information of the BLASTx sequence alignment search for the query sequence into several training data sets. The alignment and entry information is quantitatively filtered and scored. Only one entry from each contributing database per alignment is used to construct the data training set to prevent skewing the data set by over-representation of a single source. The entry with the highest annotation score is selected as the best description of the query sequence and the annotation is transferred to the query as the CAPASA function.

(i) Sequence Homology Score (η)

The BLASTx sequence similarity search of query sequence will return some total number of results (X_{tot}) from the database depending on the length and complexity of the query. The size of X_{tot} will vary from as low as zero for short or non-complex sequences that do not satisfy the BLASTx alignment parameters, to several hundred for highly conserved functions. The homology score is a single measure that accounts for the absolute sequence alignment to the query, the relative quality of the alignment relative to other alignments in the output, and the probability that the current alignment was selected

by chance from the nr sequence database. The homology score is the bit score ratio of each alignment (S_x') over the best bit score (S_{max}') multiplied by a correction factor E^*

$$\eta_x = \left(\frac{S_x'}{S_1'}\right)E^* \qquad \textbf{Eq. 5}$$

where

$$E^* = \frac{-\log_{10}(E_x)}{180} \qquad \textbf{Eq. 6}$$

The major component of the E-score is the term λ within the exponential bit score S′. By using the negative logarithm of the E-score we change the importance of the component factors such that the most influential term is now the length of the query sequence n, because the residue length of the database (m) is large enough to be effectively constant. The E* term reduces the large positive contribution to the total annotation score of strong alignment to low quality query sequences that will occur if the best sequence alignment in the BLASTx output has a low E-score. Secondly, the E* term serves to linearize the calculation of the probability that a better sequence does not exist in the database across the range of all E-scores (Figure 2).

(ii) Taxonomy Relation Score (τ)

We have developed a method of constructing a composite source organism for the query sequence based on the homology and taxonomic lineage of the entries (Y_{tot}) associated with each alignment entry as an approximation of their phylogenetic relationship. CAPASA parses the entry information for each organism (from one entry per contributing database per alignment) into an abbreviated taxonomic lineage using information from the NCBI Taxonomy Database (Benson et al. 2000; Wheeler et al. 2002) into lists of taxons for each major category ($\chi 1$ = Superkingdom, $\chi 2$ = Kingdom,

$\chi 3$ = Phylum, $\chi 4$ = Order, $\chi 5$ = Class, $\chi 6$ = Family, $\chi 7$ = Genus, $\chi 8$ = Species/strain). The value of a particular taxon in the category is its homology-weighted frequency (φ_y^{χ}), which is the average bit score for the taxon times its frequency in the taxonomic category

$$\varphi_{\chi y}^{\tau} = \left(\overline{S}_x^{'}\right)_y \text{ x } \left(\text{taxon frequency}\right)_{Ytot}^{\chi} \qquad \textbf{Eq. 7}$$

Information derived from higher homology sequence alignments should be more reliable for annotation transfer to the query sequence. The combined homology weighted frequency increases the reliability of the particular taxon measurement by measuring both the sequence confidence and its reliability based on its selection by other members of the scientific community who deposited the information in the sequence and taxonomy databases. The taxonomy score of the entry (τ_y) is one-eighth of the sum of the ratio of the entry homology weighted frequency over the best homology weighted frequency for each of the eight taxonomy categories measured with CAPASA multiplied by a correction factor

$$\tau_y = \sum_{\chi=1}^{\chi=8}\left(\frac{\varphi_{\chi y}^{\tau}}{\varphi_{\chi max}^{\tau}}\right) x \left(1-\frac{\mu_y}{\nu}\right) \qquad \textbf{Eq. 8}$$

The factor μ_y is the number of sequences present in the Genbank protein database associated with the organism y, while ν is the total number of protein sequences contained in Genbank at the time the BLASTx search is performed. If a particular organism is represented by an appreciably large number of sequences in the nr protein database the likelihood of selecting a sequence derived from that organism by chance increases. The database correction factor ($1-\mu_y/\nu$) is the probability that the organism is *not* associated with the alignment by chance. For most organisms μ_y is relatively small versus the total number of sequences in the protein database and this term is irrelevant,

however for many model organisms or species whose full genome sequence has been determined the taxonomy database correction factor is crucial to properly calculating the CAPASA taxonomy score.

(iii) Consensus Name Score (γ)

The most important feature of CAPASA is the transfer of a putative functional assignment to the query sequence. CAPASA constructs a training set for name selection based on the individual words that comprise the gene product name or function. Words that are shorter than two characters, words composed of more than 40% numbers, words that imply less certainty such as "hypothetical", and certain common language words are not included in the training set. Importantly, the words "conserved", "homolog", "probable", "putative", and "similar" are included in the construction of the training data set because they often indicate the particular entry was annotated by sequence homology comparison instead of experimental validation. The value of a particular word is the homology-weighted frequency of the word (φ_y^γ) in the training set of words

$$\varphi_y^\gamma = \left(\overline{S}_y^{'}\right)_x \text{ x } \left(\text{word frequency}\right)_{Ytot} \qquad \textbf{Eq. 9}$$

Like the taxonomy lineages, only one function name per contributing database per alignment is used to construct the training set of words. The total value of a particular function name (γ_y) is the sum of the homology weighted frequency ratio of the component word over the best homology weighted frequency normalized by the length of the name (N_y)

$$\gamma_y = \frac{\sum_N \left(\varphi_y^\gamma / \varphi_{max}^\gamma\right)}{N_y} \qquad \textbf{Eq. 10}$$

As with the taxonomy, the homology-weighted frequency ratio quantifies both reliability and popularity of the component words for each gene name. By quantifying the individual words, as opposed to whole names, it is possible for CAPASA to differentiate between minor differences in gene names, select against misspellings (Gilks et al. 2002), and select against overly specific or descriptive annotations.

CAPASA Software Validation

The performance of CAPASA annotation selection and transfer was benchmarked by comparison to several automated or manually supervised methods of annotation assignment. A variety of prokaryotic and eukaryotic sequence sources were used in the evaluation to highlight the robustness and flexibility of CAPASA to work with any sequence that can be analyzed by BLASTx search. Comparisons were made to annotations assigned by human experts against a set of expressed sequence tags (ESTs) from monocot plants (Rice Anchor Set) (Van Deynze et al. 1998), the genome of the yeast *Saccharomyces cerevisiae* annotated from literature curation by a committee of experts from the Saccharomyces Genome Database (Issel-Tarver et al. 2002; Weng et al. 2003), and a selection of sequences annotated by the GeneQuiz automated annotation system (Andrade et al. 1999; Iliopoulos et al. 2001). Lastly, the validated CAPASA program was applied to the annotation of the full *Rhodococcus* sp. I24 genome from the ERGO™ database of Integrated Genomics Inc. (Chicago, IL.) (Overbeek et al. 2003). The annotations derived from the external sources will be referred to collectively as the "expert annotations".

The accuracy of annotation assignment was determined by text string comparison of the CAPASA annotation to the expert annotation. Initial matches were determined by

computational matching (identical text string or embedded substring matches where the full text on one annotation was embedded within the other), secondary matches were assigned by manual inspection of the two annotations to determine if their biological meaning was clearly similar. A query sequence was deemed unscoreable if the "expert annotation" or CAPASA assigned annotation was noninformative. An annotation assignment was noninformative if: it was empty (expert annotations with no assignment or failed BLASTx searches), the annotation contained no information about biological function (enzyme activity, cellular phenotype, or some sort of biological function), the annotation was a database identifier (e.g. "Ydr524cp; CAI: 0.14" as a descriptor for the *S. cerevisiae* gene AGE1). Many sequences could not be processed by CAPASA because the BLASTx search failed to find any database sequences with minimal homology. The full set of genome scale annotations is available online at http://web.mit.edu/biology/sinskey/www/home.html in the publications section.

(i) Rice Anchor Set

The Rice Anchor Set is a collection of plant cDNA clones selected for their ability to hybridize to a wide variety of grass genera for comparative hybridization and genome analysis (Van Deynze et al. 1998). The collection contains ESTs selected from the agriculturally important plant species *Avena sativa* (oat), the fully sequenced *Oryza sativa* (rice), and *Hordeum vulgare*, (barley). Van Deynze et-al. (Van Deynze et al. 1998) reported annotations of the anchor set cDNAs based on manual examination of BLASTx search results. We used CAPASA to assign functional annotations to these cDNA sequences and compared our results with the previously published annotations. Only 34 of the 152 sequences listed in the paper were directly scoreable against

CAPASA as many query sequences returned no result in our BLASTx search, or because there was no descriptive annotation in the reference. 32 of the 34 (94%) scoreable annotations from the paper agreed with annotations assigned based on the sequence homology search results by CAPASA. The two scored mismatches were RZ244R, described by Van Deynze et al. (Van Deynze et al. 1998) as "ferric leghemoglobin reductase" and RZ995, "hypothetical ferripyochelin binding protein". The CAPASA annotation to RZ244R "putative dihydrolipoamide dehydrogenase precursor" matched the annotation of RZ244F (the forward sequence primer of the same clone), while the CAPASA description of RZ995 "transferase hexapeptide repeat family" had no relation to its counterpart in the literature.

(ii) Saccharomyces Genome Database

The Saccharomyces Genome Database (SGD) is one of the foremost global stores of sequence, physiology, and metabolism of the model organism *S. cerevisiae*. The information maintained in the SGD is based on continuous literature curation by a committee of experts in the yeast research community. Protein coding DNA sequences and gene annotation assignments were obtained from the SGD (http://genome-www.stanford.edu/Saccharomyces/). SGD annotations were compared to CAPASA annotations at the level of gene name (functional description) or standard name (the letter and number combination gene designation) to determine matches. 4294 of 4366 yeast ORFs contained the required information to make a scoreable comparison, of these 217 (5.1%) were scored as identical annotation, 3021 (70.4%) contained substring matches, and an additional 784 (18.3%) were assigned as matches by manual inspection. The total agreement between SGD assigned annotation and CAPASA annotations assigned from

BLASTx results was 93.7%. The majority of mismatched CAPASA assignments were annotated as various uncharacterized functions ("hypothetical protein", "unnamed protein product"), highly conserved functions (transcription regulators, ribosomal proteins), or noninformative locus assignments. The total processing time for annotation of the yeast genome was just over eight hours with the BLASTx-CAPASA combination running 11 routines in parallel (49 annotations per hour per copy of the program).

(iii) GeneQuiz

GeneQuiz is an internet analysis and viewing system for transfer of functional annotation to novel protein sequences developed by the European Bioinformatics Institute (Cambridge, UK) (Andrade et al. 1999; Iliopoulos et al. 2001). The GeneQuiz system employs modular system architecture of multiple database maintenance tools, sequence alignment tools, and lexical analyses (http://jura.ebi.ac.uk:8765/ext-genequiz/). The number of sequences compared between CAPASA and GeneQuiz was limited to 24 due to constraints imposed on throughput of the GeneQuiz email queuing system. These sequences were selected from the Rice Anchor Set and SGD yeast sequences. The Rice Anchor Set DNA sequences were translated to their protein counterpart using the EBI Translation Machine (http://www2.ebi.ac.uk/translate/) in the frame of the highest homology selection returned by the BLASTx. SGD ORF translation sequences were used as the yeast sequence dataset data set for GeneQuiz analysis. Only two annotations disagreed between the two systems, both of which were listed with marginal reliability scores by GeneQuiz, although these two CAPASA annotations agreed with the expert source annotations from SGD (TABLE 1).

(iv) ERGO™ Genome Database

Rhodococcus sp. I24 is a Gram-positive nocardioform actinomycete capable of degrading naphthalene and toluene as sole carbon sources (Chartrain et al. 1998; Treadway et al. 1999). The genome sequence was determined by Integrated Genomics Inc. and is available via the ERGO™ database at (http://ergo.integratedgenomics.com/ERGO/) (Overbeek et al. 2003). ERGO™ is a subscription accessible database, which uses genome comparison methods to determine the function of novel genes in newly sequenced or poorly characterized organisms by chromosomal synteny, metabolic pathway prediction, and missing gene analysis (Bork et al. 1998; Consortium 2001). The Integrated Genomics ERGO™ database contained 6098 open reading frame sequences for *Rhodococcus* sp. I24 in 2003, not all of which possessed an annotation assignment. A total of 3037 of these contained scoreable annotations from both ERGO™ and CAPASA for comparison. Of these 135 (4.4%) were identical text matches between ERGO™ and CAPASA annotation assignments, 896 (29.5%) were matched by text substring comparison, while 1453 (47.8%) were assigned as matches by inspection. The remaining 553 (18.2%) annotations disagreed between ERGO™ and CAPASA. A total of 3061 sequence annotations were not scoreable. Of these, 1997 ERGO™ entries had no information; 1064 annotations from either ERGO™, CAPASA, or both were noninformative. CAPASA was able to assign functional annotation to 445 of the sequences that had no information assigned from the ERGO™ database. The total processing time for annotation of the *Rhodococcus* sp. I24 genome was just over eight hours with the BLASTx-CAPASA combination running 13 routines in parallel (39.5 annotations per hour per copy of the program).

Discussion

As the speed of sequence generation increases it will be crucial for the scientific community to develop and use automated annotation assignment methodologies as a first step in creating a framework for investigating and determining the purpose and function of newly sequenced genes. CAPASA is a new method to be included in the global toolbox of annotation systems. The advantage of CAPASA over other systems is that it does not require a large amount of computing power. It is fully functional on a desktop computer running PERL in the Windows environment. Its small size (165 kb) and flexibility (DNA or protein sequences and previously generated BLAST output files can be used as input for different modules) makes CAPASA readily available to the small sequencing efforts initiated by individual researchers as well as the large-scale sequencing consortium. CAPASA is fast enough to complete genome scale projects in a matter of days while being simple enough for the average researcher to implement on a desktop computer without IT support. CAPASA is based on well-established sequence similarity search systems and objective selection rules. The annotations transferred to the query sequence are rigorously analyzed and are based on the most complete information publicly available at the time the BLAST search is executed.

The Consensus Annotation by Phylogeny Anchored Sequence Alignment algorithm for functional annotation transfer successfully assigns annotations to query sequences with good agreement to several other methods including expert selected literature curation (Saccharomyces Genome Database), manual selection from BLAST search results (Rice Anchor set), complex computational database analysis (GeneQuiz), or genome comparison methodologies (ERGO^{TM}). CAPASA performed at a level of 82-

94% agreement to these various systems at speeds that allowed for the annotation of full microbial genomes overnight. These results were achieved with a desktop computer and the popular PERL programming environment. The premise that similar sequences from similar organisms perform similar functions has successfully been captured by CAPASA. Most importantly, the annotation score of CAPASA assignment is a strong measure of alignment quality combining information about both the sequence alignment similarity as well as the combined agreement of the global research community about functional assignment and organism relatedness.

Acknowledgments

We would like to thank X.M. O'Brien, P.A. Lessard, L.B. Willis, and Prof. C. Burge for helpful suggestions. Special thanks are extended to T. Tao and S. McGinnis at NCBI for useful discussions and assistance with QBlast URL API searches and access to the taxonomy database.

List of abbreviations

S	BLAST raw homology score
λ	BLAST sequence composition metric
K	BLAST sequence distribution metric
S′	BLAST bit score
E	BLAST expect score
m	Size of Genbank database in amino acid residues
n	Size of query sequence in amino acid residues
x	Alignment number in BLASTx output
y	Entry number within a BLASTx alignment
α_{xy}	CAPASA annotation score
η_x	CAPASA homology score of the alignment 'x'
τ_y	CAPASA taxonomy score of the entry 'y'
γ_y	CAPASA consensus name score for the entry 'y'
E^*	Un-E-score, log transformation of the BLAST expect score
φ^τ	Homology weighted frequency of the taxonomy score
φ^γ	Homology weighted frequency of the consensus name score
χ	Taxonomic category
μ_y	Number of sequence entries from the organism of entry 'y' in Genbank non-

redundant protein database

ν Total number of sequences in the Genbank non-redundant protein database

Figure 1. Schematic representation of CAPASA workflow. The BLASTx search output is parsed into homology, taxonomy, and annotation components. The taxonomy lineage and functional annotation associated with each alignment entry is used to construct training sets to quantify the homology-weighted frequency of each component. These are combined to determine the annotation score for each entry. The entry with the highest annotation score is assigned as the putative function of the query sequence.

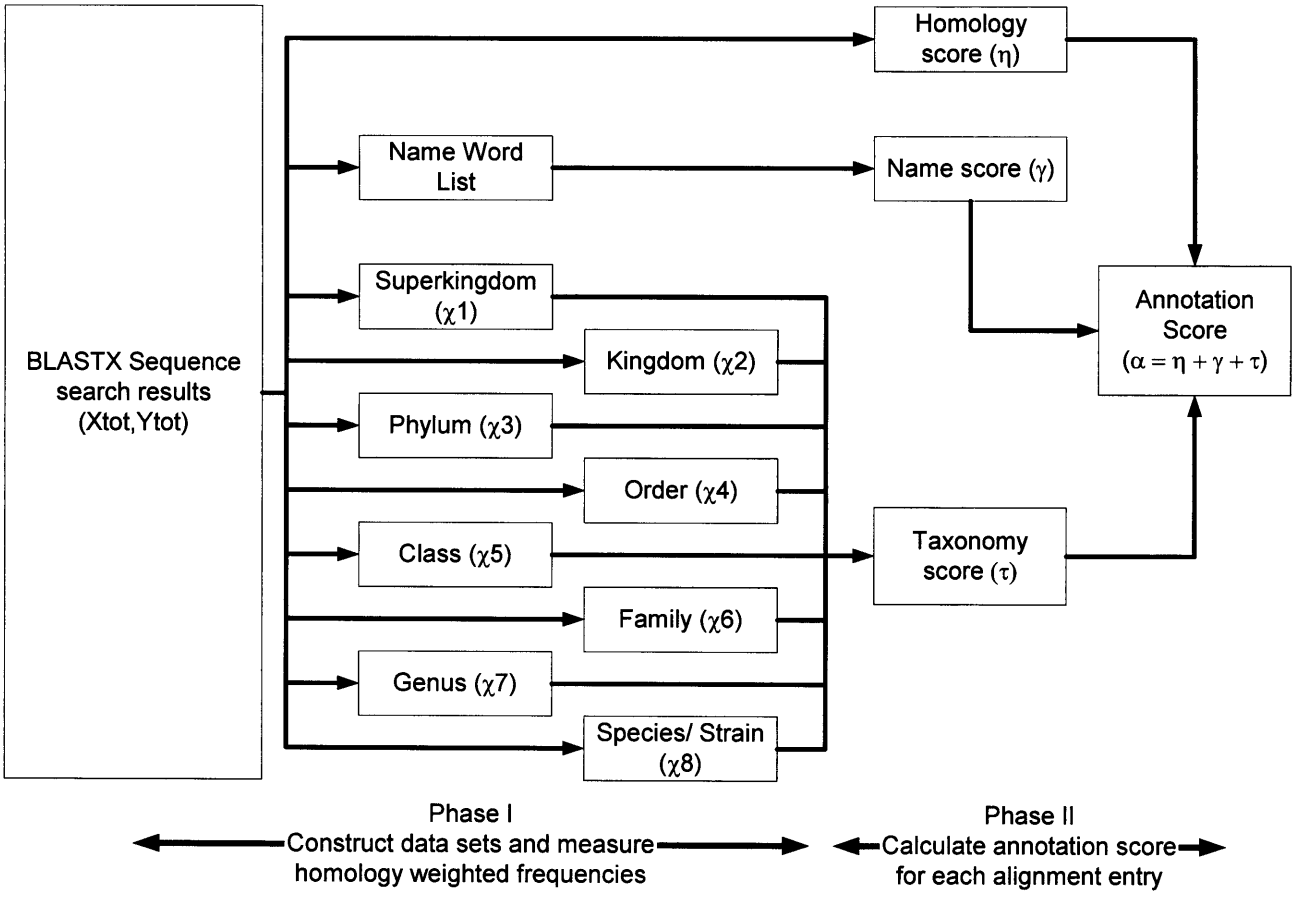


Figure 2. A comparison of probability (■) and E* (▲) versus the E-score of an alignment sequence. The probability saturates very quickly at uninformative high E-scores. E* is linear across the entire range of E-score possibilities, allowing for a fine discrimination of the quality of sequence similarity between highly homologous sequences.

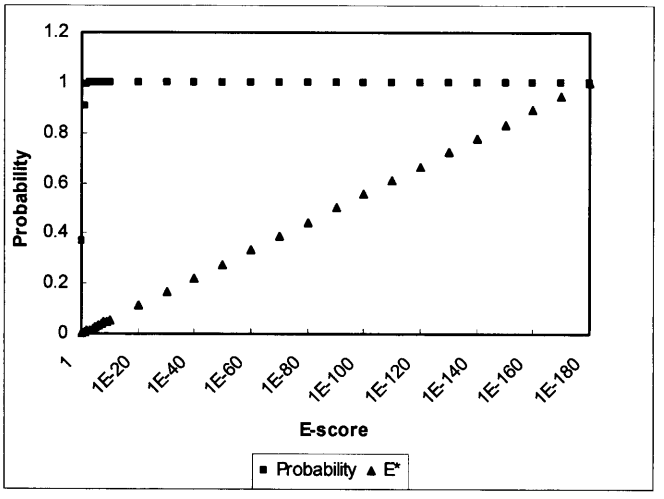


Table 1.

Sequence information			GeneQuiz		CAPASA	
ORF	Source organism	Length (aa)	GeneQuiz Annotation	Reliability Score	CAPASA annotation	Annotation Score
CDO1081F	*Avena sativa*	70	ENDO-POLYGALACTURONASE-LIKE PROTEIN	0.99 (clear)	polygalacturonase, putative	2.606
CDO202	*Avena sativa*	105	PUTATIVE CARRIER PROTEIN (AT4G01100/F2N1_16)	0.99 (clear)	similar to mitochondrial carrier family	2.433
BCD808	*Hordeum vulgare*	89	Tubulin alpha-2 chain	0.99 (clear)	alpha tubulin	2.982
BCD880	*Hordeum vulgare*	50	Glutathione reductase, cytosolic (EC 1.6.4.2	0.95 (clear)	glutathione reductase	2.039
RZ400	*Oryza sativa*	122	Ras-related protein RHN1	0.99 (clear)	GTP-binding protein (Ara6)	2.265
RZ476	*Oryza sativa*	49	Elongation factor 1-gamma (EF-1-gamma) (eEF-1B gamma)	0.99 (clear)	Elongation factor 1-gamma	2.965
RZ508	*Oryza sativa*	80	Catalase isozyme B (EC 1.11.1.6) (CAT-B)	0.99 (clear)	catalase (EC 1.11.1.6) 1	2.922
RZ567	*Oryza sativa*	64	KINESIN LIGHT CHAIN (FRAGMENT)	0.99 (clear)	putative kinesin light chain	2.472
RZ672	*Oryza sativa*	93	CYSTATHIONINE GAMMA-SYNTHASE (EC4.2.99.9) (FRAGMENT)	0.99 (clear)	cystathionine gamma-synthase	2.844
RZ900	*Oryza sativa*	120	Adenosylhomocysteinase (EC 3.3.1.1) (S-adenosyl-L-homocysteine hydrolase) (Ado	1 (clear)	adenosylhomocysteinase	2.730
AAC1	*Saccharomyces cerevisiae*	309	ADP,ATP carrier protein 1 (ADP/ATP translocase 1) (Adenine nucleotide translocat	1 (clear)	ADP/ATP translocator	2.604
M1	*Saccharomyces cerevisiae*	585	397AA LONG HYPOTHETICAL NADH OXIDASE	0.3 (marginal)	cytochrome c trimethylase; Ctm1p	2.979
GIP2	*Saccharomyces cerevisiae*	548	GLC7-interacting protein 2	1 (clear)	Gip2p	2.876
HRR25	*Saccharomyces cerevisiae*	494	Casein kinase I homolog HRR25 (EC 2.7.1.-)	1 (clear)	casein kinase I	1.981
IMP1	*Saccharomyces cerevisiae*	190	Mitochondrial inner membrane protease subunit 1 (EC 3.4.99.-)	1 (clear)	membrane protease 1	2.503
KNH1	*Saccharomyces cerevisiae*	268	Cell wall synthesis protein KNH1 precursor	1 (clear)	Knh1p	2.750
LYS20	*Saccharomyces cerevisiae*	428	Homocitrate synthase, cytosolic isozyme (EC 4.1.3.21)	1 (clear)	Homocitrate synthase, cytosolic isozyme	2.331
I	*Saccharomyces cerevisiae*	406	RESTRICTION-MODIFICATION ENZYME SUBUNIT R3 (EC 3.1.21.3)	0.3 (marginal)	Outer Kinetochore Protein; Okp1p	2.729
PZF1	*Saccharomyces cerevisiae*	429	Transcription factor IIIA (TFIIIA)	1 (clear)	Transcription factor IIIA	2.957
RPC11	*Saccharomyces cerevisiae*	110	DNA-directed RNA polymerase III 12.5 kDa polypeptide (EC 2.7.7.6)	1 (clear)	DNA-directed RNA polymerases III 12.5 kDa polypeptide	1.996
RPL37B	*Saccharomyces cerevisiae*	88	60S ribosomal protein L37-B (L35) (YP55)	0.99 (clear)	ribosomal protein L37 e.B, cytosolic	1.429
SOL4	*Saccharomyces cerevisiae*	255	Probable 6-phosphogluconolactonase 4 (EC 3 1 1.31) (6PGL)	1 (clear)	Probable 6-phosphogluconolactonase 4	2.671
SUT1	*Saccharomyces cerevisiae*	299	Probable sterol carrier	1 (clear)	Involved in sterol uptake; Sut1p	2.286

Table 1. Results of functional annotation of 24 sequences by GeneQuiz (Andrade et al. 1999) and CAPASA. The grey highlighted entries indicate entries whose annotations disagreed between the two systems. The annotations for these two genes from the Saccharomyces Genome Database are CTM1: cytochrome c methyltransferase and OKP1: outer kinetochore protein (Dwight et al. 2002).

References

Adams, MD, SE Celniker, RA Holt, CA Evans, JD Gocayne, PG Amanatides, SE Scherer, PW Li, RA Hoskins, RF Galle, et al. (2000). The genome sequence of *Drosophila melanogaster*. Science **287**(5461): 2185-95.

Altschul, SF (1991). Amino acid substitution matrices from an information theoretic perspective. J Mol Biol **219**(3): 555-65.

Altschul, SF, W Gish, W Miller, EW Myers and DJ Lipman (1990). Basic local alignment search tool. J Mol Biol **215**(3): 403-10.

Andrade, MA, NP Brown, C Leroy, S Hoersch, A de Daruvar, C Reich, A Franchini, J Tamames, A Valencia, C Ouzounis, et al. (1999). Automated genome sequence analysis and annotation. Bioinformatics **15**(5): 391-412.

Benson, DA, I Karsch-Mizrachi, DJ Lipman, J Ostell, BA Rapp and DL Wheeler (2000). GenBank. Nucleic Acids Res **28**(1): 15-8.

Benson, DA, I Karsch-Mizrachi, DJ Lipman, J Ostell and DL Wheeler (2003). GenBank. Nucleic Acids Res **31**(1): 23-7.

Blattner, FR, G Plunkett, 3rd, CA Bloch, NT Perna, V Burland, M Riley, J Collado-Vides, JD Glasner, CK Rode, GF Mayhew, et al. (1997). The complete genome sequence of *Escherichia coli* K-12. Science **277**(5331): 1453-74.

Bork, P, T Dandekar, Y Diaz-Lazcoz, F Eisenhaber, M Huynen and Y Yuan (1998). Predicting function: from genes to genomes and back. J Mol Biol **283**(4): 707-25.

Brenner, SE (1999). Errors in genome annotation. Trends Genet **15**(4): 132-3.

Chartrain, M, Jackey, B., Taylor, C., Sandford, V., Gbewonyo, K., Lister, L., Dimichele, L., Hirsch, C., Heimbuch, B., Maxwell, C., Pascoe, D., Buckland, B., Greasham, R. (1998). Bioconversion of indene to *cis* (1S,2R) indandiol and *trans* (1R,2R) indandiol by *Rhodococcus* species. Journal of Fermentation and Bioengineering **86**(6): 550-558.

Gene Ontology Consortium, (2001). Creating the gene ontology resource: design and implementation. Genome Res **11**(8): 1425-33.

Dwight, SS, MA Harris, K Dolinski, CA Ball, G Binkley, KR Christie, DG Fisk, L Issel-Tarver, M Schroeder, G Sherlock, et al. (2002). Saccharomyces Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). Nucleic Acids Res **30**(1): 69-72.

Gilks, WR, B Audit, D De Angelis, S Tsoka and CA Ouzounis (2002). Modeling the percolation of annotation errors in a database of protein sequences. Bioinformatics **18**(12): 1641-9.

Gish, W and DJ States (1993). Identification of protein coding regions by database similarity search. Nat Genet **3**(3): 266-72.

Goffeau, A, BG Barrell, H Bussey, RW Davis, B Dujon, H Feldmann, F Galibert, JD Hoheisel, C Jacq, M Johnston, et al. (1996). Life with 6000 genes. Science **274**(5287): 546, 563-7.

Iliopoulos, I, S Tsoka, MA Andrade, P Janssen, B Audit, A Tramontano, A Valencia, C Leroy, C Sander and CA Ouzounis (2001). Genome sequences and great expectations. Genome Biol **2**(1): INTERACTIONS0001.

Issel-Tarver, L, KR Christie, K Dolinski, R Andrada, R Balakrishnan, CA Ball, G Binkley, S Dong, SS Dwight, DG Fisk, et al. (2002). Saccharomyces Genome Database. Methods Enzymol **350**: 329-46.

Karlin, S and SF Altschul (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc Natl Acad Sci U S A **87**(6): 2264-8.

Koski, LB and GB Golding (2001). The closest BLAST hit is often not the nearest neighbor. J Mol Evol **52**(6): 540-2.

Lander, ES, LM Linton, B Birren, C Nusbaum, MC Zody, J Baldwin, K Devon, K Dewar, M Doyle, W FitzHugh, et al. (2001). Initial sequencing and analysis of the human genome. Nature **409**(6822): 860-921.

Overbeek, R, N Larsen, T Walunas, M D'Souza, G Pusch, E Selkov, Jr., K Liolios, V Joukov, D Kaznadzey, I Anderson, et al. (2003). The ERGO genome analysis and discovery system. Nucleic Acids Res **31**(1): 164-71.

Schloss, PD and J Handelsman (2003). Biotechnological prospects from metagenomics. Curr Opin Biotechnol **14**(3): 303-10.

Treadway, SL, KS Yanagimachi, E Lankenau, PA Lessard, G Stephanopoulos and AJ Sinskey (1999). Isolation and characterization of indene bioconversion genes from *Rhodococcus* strain I24. Appl Microbiol Biotechnol **51**(6): 786-93.

Van Deynze, AE, ME Sorrells, WD Park, NM Ayres, H Fu, SW Cartinhour, E Paul and SR McCouch (1998). Anchor probes for comparative mapping of grass genera. Theoretical and Applied Genetics **97**(3): 356-369.

Venter, JC, MD Adams, EW Myers, PW Li, RJ Mural, GG Sutton, HO Smith, M Yandell, CA Evans, RA Holt, et al. (2001). The sequence of the human genome. Science **291**(5507): 1304-51.

Wall, L, Christiansen, T., Orwant, J. (2000). Programming Perl. Sebastopol, CA, O'Reilly & Associates, Inc.

Weng, S, Q Dong, R Balakrishnan, K Christie, M Costanzo, K Dolinski, SS Dwight, S Engel, DG Fisk, E Hong, et al. (2003). Saccharomyces Genome Database (SGD) provides biochemical and structural information for budding yeast proteins. Nucleic Acids Res **31**(1): 216-8.

Wheeler, DL, DM Church, AE Lash, DD Leipe, TL Madden, JU Pontius, GD Schuler, LM Schriml, TA Tatusova, L Wagner, et al. (2002). Database resources of the National Center for Biotechnology Information: 2002 update. Nucleic Acids Res **30**(1): 13-6.

Yu, J, S Hu, J Wang, GK Wong, S Li, B Liu, Y Deng, L Dai, Y Zhou, X Zhang, et al. (2002). A draft sequence of the rice genome (*Oryza sativa L. ssp. indica*). Science **296**(5565): 79-92.

Chapter III

Trigonometric deconvolution analysis of DNA microarrays from *Rhodococcus* sp. I24 aromatic hydrocarbon fermentations.

Abstract

The genus *Rhodococcus* is gaining value as a versatile platform for biocatalytic manufacturing of chiral intermediates and small molecules (O'Brien et al. 2002). *Rhodococcus* sp. I24 was grown by batch fermentation in the presence of naphthalene and indene to measure changes in gene expression and aromatic hydrocarbon metabolism with DNA microarray technology. We describe the theory and application of trigonometric deconvolution analysis for the measurement of changes in gene expression across multiple growth conditions with a minimal number of labeled cDNA hybridizations. The analysis of gene expression and metabolite synthesis from different aromatic substrates across mid-log and late stationary growth indicates that genes associated with hydrocarbon metabolism are regulated by substrate specific and growth phase dependent mechanisms.

Introduction

The genus *Rhodococcus* has shown increasing utility as a platform system for fine chemical manufacturing with biocatalytic processes. Multi-ton production of acrylamide (Yamada et al. 1996), environmental remediation of halocarbons (Swanson 1999), and production of chiral synthons (Orru et al. 1999; O'Brien et al. 2002) represent just a few of the activities available in the metabolic repertoire of these organisms (Warhurst et al. 1994). The study of this group is complicated by recent reclassification as a genus distinct from the closely related Corynebacteria, Mycobacteria, and Nocardia (Bell et al. 1998). There have been regular fluctuations in classification criteria (Goodfellow et al. 1998) and only limited tools for modification and manipulation of genetic systems

(Larkin et al. 1998). Novel genomic analysis tools may enable solutions to overcome the numerous challenges to classical biological characterization of Rhodococci.

Past studies by our research group have explored the use of *Rhodococcus* sp. I24 for the bioconversion of indene to 2R-indandiol, a critical precursor for the HIV protease inhibitor CRIXIVAN™ (Vacca et al. 1994; Reider 1997). Current manufacturing methods for Crixivan™ production rely on an expensive stereospecific manganese salen catalyst for the production of (-)cis-(1S,2R)-1-aminoindan-2-ol, or (-)-CAI (Senanayake et al. 1996; Hughes et al. 1997). Our long-term goal has been the genetic engineering of a biological system for the production of 2R indandiol for incorporation into the CRIXIVAN™ production process. Transcriptional regulation (Chartrain et al. 1998), cloning studies (Treadway et al. 1999), and metabolic flux analysis (Yanagimachi et al. 2001) have revealed the presence of multiple competing or non-productive pathways acting in indene metabolism (Figure 1). These same metabolic pathways allow *Rhodococcus* sp. I24 to consume other polycyclic aromatic hydrocarbons (PAHs) such as naphthalene and toluene as a sole carbon source.

Our current study describes efforts to correlate well-characterized aromatic hydrocarbon metabolism with data from DNA microarrays to dissect the interplay between biocatalysis and gene expression. Detailed analysis of aromatic metabolite production during batch fermentation, measurement of glucose co-utilization, and the innovation of trigonometric deconvolution analysis of DNA microarray data begin to reveal a more complete picture of the complex network of chemistry and biology at work within this system.

Materials and Methods

Fermentation

Fermentation inocula were precultured in LB medium (Difco, Detroit, MI) overnight. Fermentation cultures were grown in a defined medium as described (Stafford et al. 2002) without MOPS buffer and with 0.25 mL P2000 polypropylene glycol antifoam (Sigma-Aldrich, St. Louis, MO) per L culture in an Infors Sixfors (Appropriate Technical Resources, Laurel, MD) six-vessel fermentation system. 20 mL of inoculum was sterilely injected into each fermenter containing 500 mL gas- and temperature-equilibrated defined medium. Each culture was independently controlled at 30°C with 1000± 5 RPM agitation. The pH was maintained at 7.0 ± 0.1 by addition of 2 M NaOH or 2 M HCl (Mallinkrodt, Paris, KY). Oxygen tension was maintained at saturation by an internal feedback controlled system with 50% O_2/N_2 mixture (BOC Gases, Cambridge, MA). Aromatic hydrocarbon feeding was initiated at an OD_{600} of 2. Indene (Sigma-Aldrich) was added to the system with a 100 mL/ min filtered nitrogen stream (BOC Gases). Naphthalene (Sigma-Aldrich) was added to the system as solid flakes. All gas tubing and fittings were composed of polytetrafluoroethylene (PTFE) (Cole Parmer, Vernon Hills, IL). 250 OD units of cells were harvested around mid log phase at an OD_{600} of 5 and at the beginning of stationary growth, flash frozen in liquid nitrogen, and stored at -80°C until needed for RNA isolation. All fermentations were performed in duplicate.

High Performance Liquid Chromatography

1 mL samples from cultured cells were extracted for aromatic hydrocarbon analysis with 1 mL of 50% (v/v) isopropanol-acteonitrile (Mallinkrodt) and cleared by

centrifugation at 13,800 *g*. Cleared samples were filtered through a 0.22 polyvinylidene difluoride (PVDF) syringe filter (Alltech, Deerfield, IL.) into glass vials. Indene and naphthalene metabolite concentrations were measured by reverse phase HPLC with an Agilent Zorbax RX-C8 column coupled to a Hewlett Packard (Palo Alto, CA.) series 1050 UV detector. Analysis was performed as previously described (Treadway et al. 1999; Stafford et al. 2002).

Glucose determination was performed by HPLC analysis of aqueous extracted media samples as previously described by Guillouet et al. (Guillouet et al. 1999) with an Aminex HPX-87H column (Bio-Rad, Hercules, CA.) coupled to an Agilent Series 1100 refractive index detector.

RNA Purification

Frozen bacterial cells were thawed on ice, centrifuged for 3 min at 4105 *g* at 4°C, and suspended in 11 mL ice cold RLT buffer (RNEasy Midi-kit; Qiagen, Valencia, CA) containing β-mercaptoethanol (Sigma-Aldrich). Cells were mixed with 12 mL cold 0.1 mm Zirconia-Silica beads (Biospec Products, Bartlesville, OK) and lysed at 0°C in a Biospec Products Bead Beater (model 1107900) with six 30 sec pulse / 30 sec pause cycles. RNA was purified using an RNEasy Midi Kit (Qiagen) following manufacturer's instructions. Residual genomic DNA was removed by 15 minute incubation with DNAse I (Qiagen). RNA concentration and integrity were measured at the MIT BioMicro Center (Cambridge, MA) with an Agilent 2100 Bioanalyzer by microfluidic electrophoresis.

Oligo probe design and selection

DNA microarray oligo probes were designed against the genome of *Rhodococcus* sp. I24 with the ArrayOligoSelector program (Bozdech et al. 2003) with oligo length of 60

and target GC content of 73%. Genome sequencing was performed by Integrated Genomics Inc. (Chicago, IL). Probes chosen for inclusion in DNA microarray printing were selected based on known or suspected participation in aromatic hydrocarbon metabolism pathways (Table 1)(Treadway et al. 1999). Each of the selected sequences was obtained from MWG Biotech (High Point, NC) and Proligo (Boulder, CO), except for sequences from contigs 2214, 2224, 2226, and 2247 which were procured from Proligo only.

Microarray printing and blocking

Oligo DNA microarrays were printed with a BioRobotics (Huntingdon, UK) MicroGrid TAS printing robot with a sixteen pin print head and BioRobotics MicroSpot 2500 quill pins. Arrays were printed onto Corning Life Sciences (Acton, MA) GAPS-2 or ULTRA slides, Full Moon Biosystems (Sunnyvale, CA) cDNA or PowerMatrix, and SCHOTT Nexterion AG (Mainz, Germany) Nexterion slides. The oligo array was printed in duplicate per slide, each spot printed in sextuplet within the array. DNA oligos were resuspended in an aqueous solution of 50% (v/v) dimethylsulfoxide (DMSO, Sigma-Aldrich), or 150 mM sodium phosphate (Mallinkdrodt) buffer (pH 8.5). Slides were blocked according to manufacturer's suggestions.

Preparation of labeled cDNA

Equal microgram quantities of intact sample RNAs were combined into a common reference sample ("reference RNA") to serve as a hybridization competitor for the individual condition RNA samples ("experimental RNA"). Experimental and control RNA samples were labeled with Cy3-dUTP or Cy5-dUTP (NEN, Boston, MA) essentially as described in Loos et al. (Loos et al. 2001). Each reaction contained 25 μg

Rhodococcus RNA. 5 ng of control *Arabidopsis thaliana* RNA (Spot Report 3 kit, Stratagene) corresponding to control spots printed on the microarray was also included in the labeling reaction. After labeling, the RNA template was destroyed by alkaline lysis treatment with 0.1 M NaOH (Sigma-Aldrich) at 65°C for 15 min. Unincorporated dye and enzyme removal and buffer exchange was accomplished using a QIAquick PCR purification kit (Qiagen) following manufacturer's instructions, with the final elution performed with MilliQ water (pH 8.0). Labeled cDNA samples were dried under vacuum before hybridization.

Microarray hybridization and scanning

Hybridization and post processing were performed as described in (Loos et al. 2001) using a Corning hybridization chamber (catalog number 2551). All hybridizations were performed as "dye swapped" pairs. That is, one array on each slide was hybridized with experimental RNA labeled with Cy3-dUTP and reference RNA labeled with Cy5-dUTP, and the other array on the same slide was hybridized with Cy5-dUTP labeled experimental RNA and Cy3-dUTP labeled reference RNA. Microarrays were analyzed using an ArrayWoRx E CCD scanner (Applied Precision, Issaquah, WA). The Cy3 color channel was scanned with a 0.1 sec exposure time whereas Cy5 was scanned with a 0.3 sec exposure time. Images of each fluorescence channel were exported as 16 bit grayscale TIFF images. Spot detection and fluorescence intensity measurements were made with the MolecularWare Digital Genome software package (Cambridge, MA) with contour intensity calculation shape, annulus background calculation with 15 percent inner diameter, 85 percent outer diameter, and no ratio normalization. The total signal intensity (x_s), signal and background number of pixels (N_s and N_b respectively), mean background

intensity ($\bar{x}_b$), and signal and background standard deviation per pixel (σ_s and σ_b respectively) were exported to Microsoft Excel (Redmond, WA) for further analysis.

Microarray analysis

Statistical normalization was used to calculate the channel combined weighted average pixel intensity for each labeled cDNA (Brown et al. 2001; Loos et al. 2001). The standard deviation for each measurement was maintained through each calculation using standard methods of error propagation (Taylor 1997). All spots with a total signal-to-noise ratio less than one and genes represented by fewer than two spots on a single array were removed from analysis. References to a particular "gene" are meant to include those microarray spots with a common oligo probe sequence, oligo manufacturer, and print buffer.

(i) Normalized weighted average pixel intensity: The mean pixel signal and background intensities and the standard deviations of the mean for each spot in each array on a given slide were calculated from the measurements exported from the MolecularWare image analysis software as:

$$\bar{x}_s = \frac{x_s}{N_s} \qquad \sigma_{\bar{x}_s} = \frac{\sigma_s}{\sqrt{N_s}}$$
$$\bar{x}_b = \frac{x_b}{N_b} \qquad \sigma_{\bar{x}_b} = \frac{\sigma_b}{\sqrt{N_b}} \qquad \textbf{Eq. 1}$$

Mean background subtraction was applied to give the background subtracted average pixel intensity (BSAPI), which was used to normalize the total channel intensity of the Cy3 and Cy5 channels for each array on the slide. The scanner normalization factor (SNF) equalizes the scale of fluorescence intensity within an array of the slide such that

$$\frac{\sum \bar{x}_{s\text{-}b}^{Cy3} / SNF}{\sum \bar{x}_{s\text{-}b}^{Cy5} (SNF)} = 1 \qquad \textbf{Eq. 2}$$

Solving Eq. 3 gives the value of the SNF:

$$SNF = \sqrt{\frac{\sum \bar{x}_{s-b}^{Cy3}}{\sum x_{s-b}^{Cy5}}} \qquad \sigma_{SNF} = 0.5 \sqrt{\left(\frac{\sigma_{Cy3}}{\sum_{Cy3} \bar{x}_{s-b}} \right)^2 + \left(\frac{\sigma_{Cy5}}{\sum_{Cy5} \bar{x}_{s-b}} \right)^2} \qquad \textbf{Eq. 3}$$

The scanner normalized intensities were log transformed and expressed as a fraction of the sum of the log transformed intensity for the channel of each array. The intensity normalized average pixel intensity for each gene's spots were combined into a single weighted average measurement for each cDNA as the combined values for the labeled cDNA intensity from both arrays of a slide (Loos et al. 2001).

$$\mu_x = \frac{\sum_{Cy3+Cy5} \bar{x}_{s-b}^{SNF} / \sigma^2_{\bar{x}_{s-b}^{SNF}}}{\sum_{Cy3+Cy5} 1/\sigma^2_{\bar{x}_{s-b}^{SNF}}} \qquad \sigma_{\mu_x} = \sqrt{\frac{1}{\sum_{Cy3+Cy5} \left(1/\sigma^2_{\bar{x}_{s-b}^{SNF}} \right)}} \qquad \textbf{Eq. 4}$$

Genes with weighted average coefficient of variance greater than 10% were not included in trigonometric deconvolution analysis.

(ii) Trigonometric deconvolution: The weighted average combined fluorescence intensity of a labeled cDNA to a gene probe from both arrays of a slide represents the fraction of bound cDNA expressed by the gene. Considering the experimental and reference cDNA separately, we were able to compare changes in transcription between two physiological conditions (growth phase or aromatic substrate) by a two-dimensional graphic system with axes of reference and experimental cDNA intensity (Figure 2). The weighted average experimental and reference cDNA intensities for a given gene describe its position in the plane of transcription as the coordinate point (ref1, exp1) for the first

physiological state, and (ref2, exp 2) for the second. Trigonometry was used to derive the percent change in transcription between growth conditions as the distance between the intensity points perpendicular to the normal, or:

$$\Delta\mu_x = \frac{\sqrt{2}}{2}(\exp 2 - \exp 1) - \frac{\sqrt{2}}{2}(\text{ref } 2 - \text{ref } 1) \quad \textbf{Eq. 5}$$

It is important to only consider the change in transcription perpendicular to the normal because fluorescence intensity ratios along the normal are equal.

Results

Fermentation

We grew several 500 mL cultures of *Rhodococcus* sp. I24 in the presence or absence of indene or naphthalene to measure the effects of different aromatic hydrocarbons on gene expression. Growth of *Rhodococcus* sp. I24 cultures was monitored by hourly measurement of OD_{600} and bi-hourly sampling for glucose determination. Aromatic hydrocarbon was introduced into the system when the culture reached an OD_{600} of about 2. Indene was fed by way of a secondary gas feed with nitrogen at 100 mL/ min. About 1 g of naphthalene was added as solid flakes; crystals persisted in the culture medium until completion ensuring a saturating amount of naphthalene through the duration. Both naphthalene and no aromatic (non-induced) cultures were treated with 100 mL/ min nitrogen to replicate the secondary gas feed used with the indene grown cultures. The correlation of OD_{600}, glucose consumption, and aromatic hydrocarbon metabolism for a representative culture with each aromatic hydrocarbon is shown in FIGURE 3. The maximum specific growth rate was 0.157 ± 0.004/h for non-induced cultures, 0.074 ± 0.002/h naphthalene cultures, and 0.160 ± 0.01/h for indene cultures. The naphthalene

specific growth rate is most probably lowered by the five fold higher concentration of aromatic hydrocarbon consistently present in the culture medium.

Aromatic hydrocarbon metabolism analysis

Polycyclic aromatic hydrocarbon dioxygenase expression in *Rhodococcus* sp. I24 has been shown to be differentially regulated by different substrates (Chartrain et al. 1998); (Treadway et al. 1999). Aromatic hydrocarbon metabolite samples were measured every hour until the end of the fermentation. The metabolite profile of each substrate and the co-consumption of glucose are shown in FIGURE 3. *Rhodococcus* sp. I24 can consume naphthalene completely as a sole carbon source (Chartrain et al. 1998; Treadway et al. 1999), and these same pathways are partially responsible for the metabolism of indene. The differential regulation of aromatic metabolism activities is most apparent in the indene grown cultures (Figure 3c). 1-indenol is first detected about four hours after the indene nitrogen feed is started, followed three hours later by cis-indandiol, and three hours later by trans-indandiol. The separation of different metabolites is indicative that at least three independent activities are responsible for indene metabolism.

RNA analysis

All RNA samples were processed with an Agilent 2100 Bioanalyzer to measure concentration and sample integrity. As shown in Figure 4, all samples except number 10 (F5 ind-2) were isolated with minimal degradation of ribosomal RNA, with average yields of 0.35 μg/μL. Aliquots from all RNA samples except number 10 were pooled for use as the reference RNA for microarray hybridizations.

Microarray analysis

RNA labeling and hybridization to the spotted DNA oligos was measured by electronic scanning of fluorescence intensity. Intensity measurements of the grayscale TIFF images were used to determine the amount of labeled experimental and reference cDNA bound to each gene spot on the array. Changes in gene expression were measured by trigonometric deconvolution to determine the distance between gene expression levels of two growth conditions perpendicular to the normal. We developed this analysis method to enable the measurement of gene expression changes between two experimental conditions without comparing them directly to each other by comparison to a common reference sample. Because samples do not have to be compared directly in a pairwise fashion the total number of hybridizations is reduced, data from multiple hybridizations of common conditions can be combined to refine expression measurements, and the total standard error of measurements is minimized. Gene expression changes were calculated by comparing transcription levels on growth substrate (non-induced vs. aromatic) at mid-log and early stationary growth phase, as shown in Figure 6. Aromatic metabolism activity increases in later stages of growth, as indicated by the HPLC measurement of aromatic metabolites of naphthalene and indene. The late growth phase increase in metabolism is reflected by changes in transcription of certain genes associated with aromatic hydrocarbon metabolism. Changes in gene expression are described as the percent change in expression from the non-induced cultures to the aromatic-induced cultures as measured by trigonometric deconvolution analysis.

Non-induced cultures were compared to naphthalene-induced cultures at mid-log and early stationary growth. Cultures at mid-log show a down regulation of the putative *tid*

genes (-29.4% for *tid*AB and -14.4% for *tid*C) believed to be responsible for synthesis and dehydrogenation of cis-(1S,2R)-indandiol (Priefert, manuscript in preparation). The *tid* gene cluster is still down regulated in early stationary phase (-16.4% for *tid*AB and -8.9% for *tid*C), however not as strongly. There were no significant changes in expression of *nim*AB when comparing non-induced and naphthalene induced genes at mid-log (+4.4%) and early stationary (+6.0%) growth phase. The *nid*AB and *nid*C genes responsible for synthesis and dehydrogenation of cis-(1R,2S)-indandiol (Treadway et al. 1999) were down regulated at mid-log (-12.2% and -4.3% respectively), but up regulated in early stationary (+6.3% and +7.7%). Uncharacterized genes in the 2214, 2226, and 2247 clusters identified as aromatic dioxygenases from gene homology show +18.8%, +28.3% and -20.9% differences respectively in expression from non-induced to naphthalene induced at mid-log growth. The same genes show average expression of +21.7%, +23.3% and +1.6% differences respectively at late stationary growth, suggesting that only the 2247 gene cluster had any significant change under naphthalene growth across the period of fermentation culture.

Comparison of gene expression between non-induced and indene grown cultures at mid-log and early stationary phases of growth revealed some differences in expression patterns from the non-induced versus naphthalene induced culture comparisons (Figure 6b and 6d). As with the naphthalene grown cultures, genes of the *tid* group are strongly down regulated in indene grown cultures when compared to the non-induced cultures at mid-log growth phase (-39.9% and -23.4% for *tid*AB and *tid*C respectively) and late stationary growth phase (-25.4% and -19.2% for *tid*AB and *tid*C respectively). The putative monooxygenase *nim*AB shows significant up regulation at mid log (+19.8% over

non-induced), which increases to +43.0% at late stationary. Also, the naphthalene inducible dioxygenase genes *nid*AB and *nid*C were up regulated +15.3% and +18.0% respectively at mid-log, and went up to +41.8% and +37.7% respectively at the late stationary phase of growth. The uncharacterized 2214, 2226 and 2247 gene clusters showed average changes in gene expression of +15.5%, +21.5%, and +11.7% respectively at mid-log growth. The same genes were up regulated to +29.7%, +33.8% and +32.5% in early stationary growth.

Discussion

Rhodococcus sp. I24 was grown in batch fermentation under non-induced or aromatic induced conditions. RNA samples from mid-log and early stationary phase were labeled and hybridized to targeted DNA microarrays to measure the expression of several genes believed to participate in polycyclic aromatic hydrocarbon metabolism. Analysis of fermentation cultures by HPLC measurement of aromatic metabolites revealed an increase in metabolic activity during later stages of culture, which was reflected in changes in gene expression. Statistical normalization and trigonometric deconvolution were applied to pairwise hybridizations between growth conditions to analyze patterns of temporal and substrate induced gene expression changes (Figure 6). The pattern of gene expression suggest that transcription of the aromatic hydrocarbon metabolism pathways are regulated by particular substrates as well as physiological conditions associated with growth phase.

The *tid*-associated genes are significantly down regulated under all conditions, both naphthalene and indene substrates as well as mid-log and early stationary phases of growth, relative to expression levels of non-induced cultures. However, the expression of

the *tid* genes increases about 10% under both aromatic substrate inductions as the culture continues from mid-log to early stationary growth. These results would suggest that the *tid* genes are transcriptionally repressed by an element which is partially responsive to growth phase dependent signals, but neither indene nor naphthalene function as an activating substrate.

Alternatively, the *nid* gene cluster is strongly repressed at mid-log when grown with naphthalene, but up-regulated 15% over non-induced with indene at the same stage of growth. The expression of the *nid* dioxygenase is increased in early stationary when grown with both naphthalene and indene by about 20% over mid-log. The results suggest the *nid* gene repressor requires two signals for expression to be activated. High-level expression is activated in late stages of growth, possibly by some sort of quorum sensing signal. Aromatic specific expression is indicated by the up-regulation in mid-log on indene, but repression at the same growth phase with naphthalene. The differential pattern of expression at mid-log would indicate that indene is able to de-repress the *nid* regulator. The most likely substrate giving rise to this de-repression is the monooxygenated 1-indenol or 1-indanone, the first substrate activity arising from indene. Unfortunately the pattern of sampling cannot resolve the inducing substrates responsible for changes in other indene metabolite production.

The *nim* genes associated with a putative epoxidation reaction is specifically responsive to indene. The naphthalene grown cultures showed consistently low expression around 4-6% over non-induced cultures at both mid-log and early stationary growth phases. However, the indene grown cultures showed expression about 20% over non-induced at mid-log and over 40% above non-induced at late stationary. The

epoxidation reaction has been shown to be the source of trans-(1R,2R)-indandiol by spontaneous hydrolysis (Stafford et al. 2002), suggesting that its expression begins several hours before the onset of stationary phase when the trans-indandiol is first produced.

The scientific community has been investigating DNA microarrays for several years, and there are almost as many methods of data analysis available as there are methods of designing DNA microarrays (Brody et al. 2002; Pan 2002; Quackenbush 2002; Datta 2003; Park et al. 2003). The log 2 normalized Cy5/Cy3 ratio presents a useful description of the fold change in expression of a single gene relative to two physiological conditions, and this is completely adequate to describe the change between those two conditions (Brown et al. 2001). However, the fold change in expression becomes less useful when trying to compare more than two conditions directly. In fact, it is useless as anything other than a qualitative means of showing stronger up or down regulation across multiple conditions such as a time course or series of substrates. The fluorescence intensity ratio, or the more accurately described relative fraction of bound cDNA, is constant along the normal when data has been log transformed. If this line is treated as the actual measure of gene expression it is possible to measure the actual percent change in gene expression between any two physiological conditions, as long as each sample RNA is hybridized against a common reference sample. The methods presented in this work accomplish just such a comparison with a reference sample composed of a mixture of experimental condition RNAs from different time points in the growth phase (mid-log and early stationary) of cultures grown with multiple aromatic hydrocarbon substrates (none, naphthalene, or indene). However, the reference sample can be composed of any

other nucleic acid composition that contains a representative of each probe on the microarray (Dudley et al. 2002). Secondly, we propose that the measurement of single channel fluorescence intensity on a pixel-by-pixel basis is a more robust presentation of DNA microarray measurements. The division of the DNA microarray "spot" into hundreds of pixel measurements effectively increases the number of data points available for analysis. The resulting amplification in the number of measurements allows the calculation of fluorescence intensity as the weighted average of pixel intensities across arrays and even across slides of replicate hybridizations with robust error propagation throughout subsequent calculations, and error reduction through application of the standard deviation of the mean (Eq. 1)(Taylor 1997).

Statistical normalization transforms all data points of a single hybridization experiments into a common form across all experiments involving the particular reference sample. Ideally, the total amount of cDNA from the experimental and reference conditions loaded onto each array are the same. When the same quantity of cDNA is applied to both arrays of a single slide the total cDNA bound to each array will also be the same. The total fluorescence within each array was within 4% of equality for Full Moon Biosystems cDNA slides, and 19% for Corning GAPS2 slides. The other slide types used in this investigation had larger discrepancies, on the order of 50-60%. The equivalence of total hybridization intensity between intra-slide arrays should increase as the coverage of the target genome and overall RNA quality increase. Therefore, the sum of fluorescence of the cDNAs within both arrays should be equal, as is calculated by our normalization methods. As shown in Figure 5a, the distribution of raw fluorescence values from the two arrays can vary widely within a slide, but they do cluster along

similar trends within an array. Scanner normalization of the individual color channels (Cy3 and Cy5) within each array (A and B) reduces the effects of dye bias during the enzymatic labeling reaction, differential scanner sensitivity for the two dyes, and localized differences in DNA concentration across the array (Figure 5b). The log transformation of a data set is a standard method of making it both linear and normal, and thereby amenable to statistical analysis (Figure 5c)(Sokal 1987). Lastly, the scanner normalized log transformed data set is expressed as the fraction of the total intensity per color channel in the array to transform the fluorescence intensity data back into an expression of cDNA bound per spot (Figure 5d). The gene expression of two experimental conditions can be plotted as single points in a two dimensional system with axes of experimental hybridization and reference hybridization (Figure 2, pink and blue respectively). The total change in gene expression between the two conditions is the distance between the lines parallel to the normal representing the ratio of experimental to reference cDNA hybridization. Specifically, the change in experimental cDNA hybridization is the altitude of the right triangle with sides equal to the difference in experimental cDNA binding, while the change in reference cDNA binding is similarly the altitude of the right triangle with sides equal to the change in reference cDNA hybridization. Using standard trigonometric methods, the total change is calculated by Eq. 5. As an added benefit, the same formula is able to indicate the direction of the change in gene expression as upregulated or downregulated by the sign of the value of the change being positive or negative respectively. The formula presented in Eq. 5 measures both the absolute percent change in gene expression between the two conditions

in question, as well as indicates the up- or downregulation of that change in a single value.

The genetic picture of aromatic hydrocarbon metabolism by *Rhodococcus* sp. I24 is complicated by the presence of multiple pathways, aromatic and growth phase dependent regulatory mechanisms, and highly variable morpho-physiological growth (Finnerty 1992; Bell et al. 1998; Goodfellow et al. 1998; Larkin et al. 1998; Treadway et al. 1999). Future developments such as full genome microarrays, high density time point experiments (RNA sampling every 15-30 minutes), and a full survey of aromatic substrates metabolized by this strain will aid in the resolution of this problem. An understanding of which genes are responsible for indene metabolism at the various time points when individual activities are induced will allow the isolation and full characterization of the exact genetic pathways responsible for production of 2R-indandiols and the development of a production scale biocatalytic system.

Acknowledgements

We would like to thank L.B. Willis for helpful suggestions and critical review in the preparation of this manuscript. This work was supported in part by a grant from E.I. DuPont and Company.

Figure 1. *Rhodococcus* sp. I24 is able to metabolize indene to multiple products including cis-(IS, 2R)-indandiol and trans-(1R, 2R)-indandiol, both of which can serve as precursors for cis-(1S)-amino-(2R)-indanol (modified from Treadway et al. 1999).

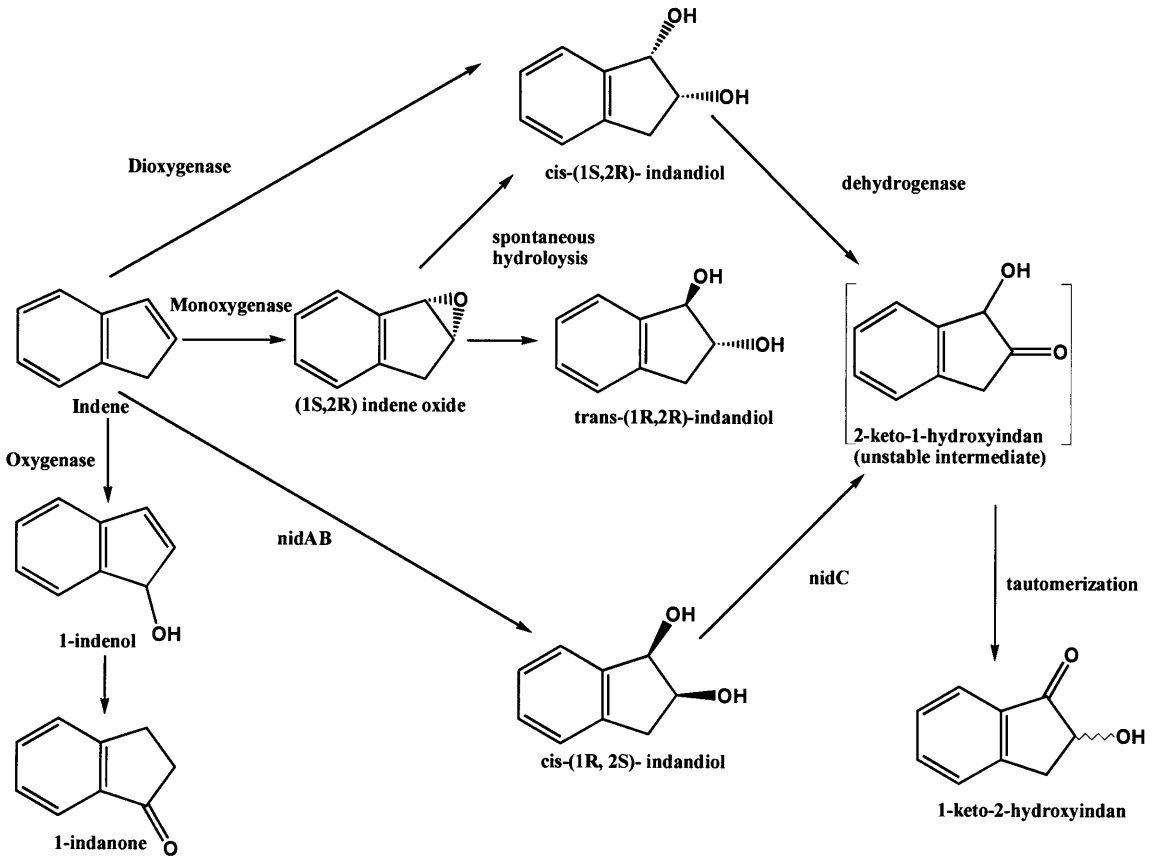


Figure 2. Trigonometric deconvolution. Changes in gene expression between two physiological conditions are described by the point coordinates of the intensity of the experimental and reference cDNAs of the gene. The change in expression is the distance between the two points perpendicular to the normal, as calculated using standard trigonometric methodologies. Background spots are plotted for explanatory value.

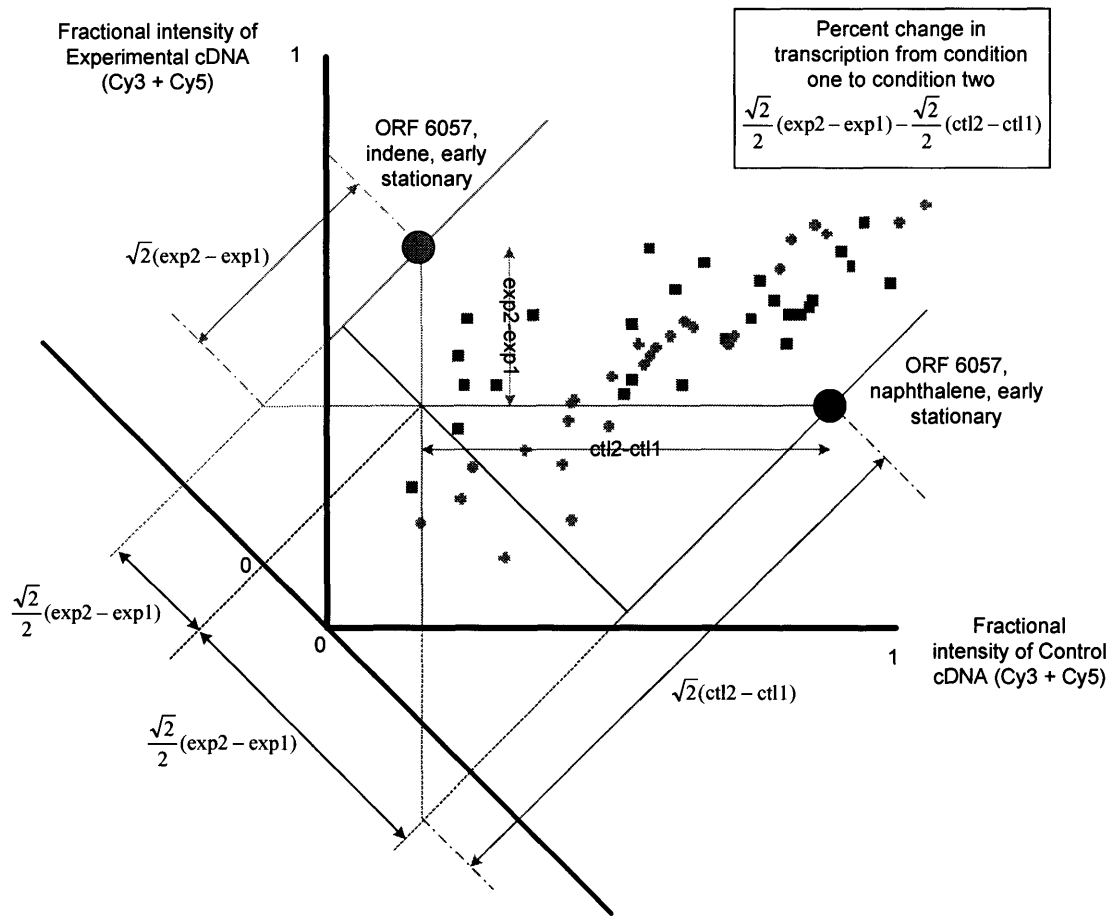
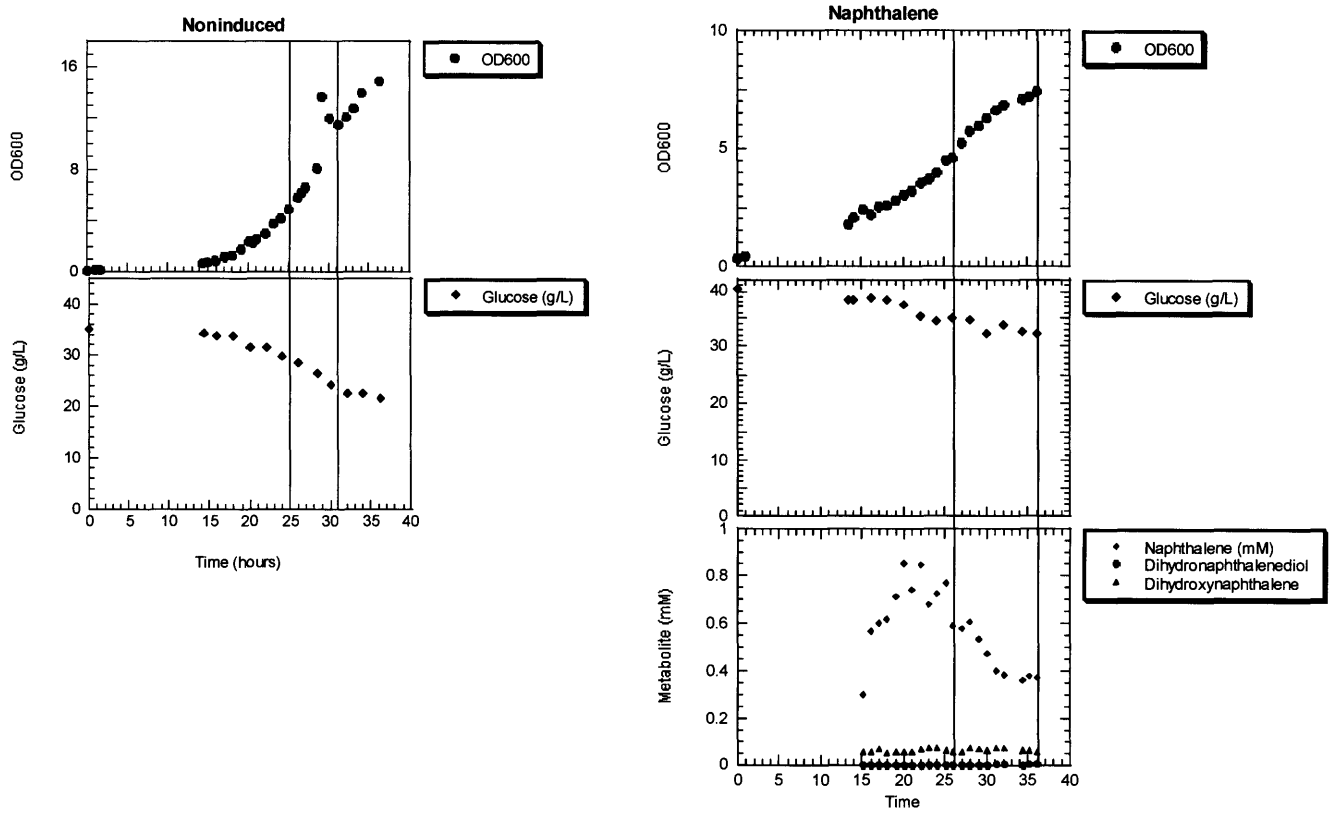


Figure 3. Fermentation of aromatic hydrocarbons. *Rhodococcus* sp. I24 fermentation cultures were monitored for about 35 hours of growth. Vertical lines through each graph indicate time points where RNA samples were harvested for DNA microarray analysis. Each graph is from measurements of a single fermentation; although all fermentations were performed in duplicate.



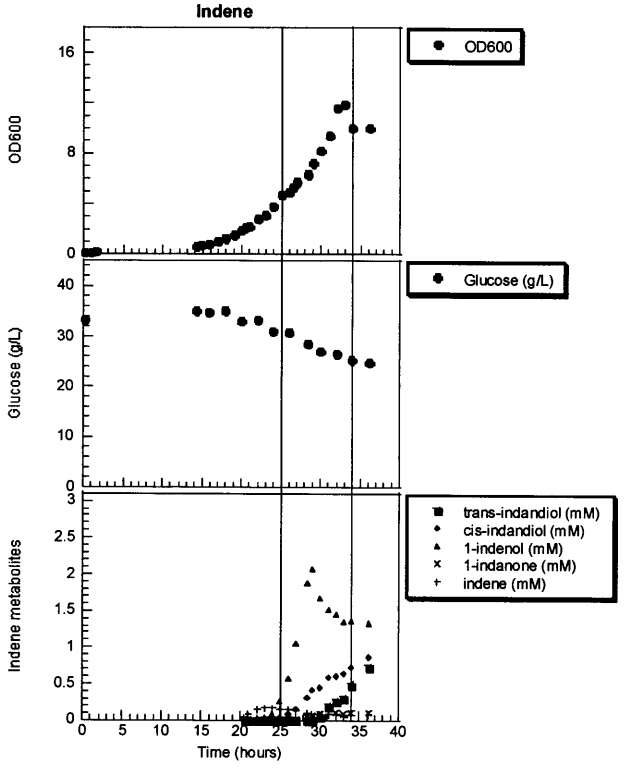
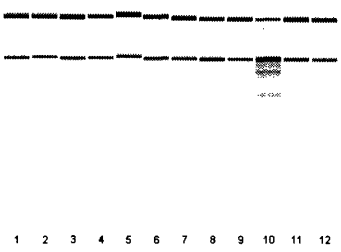

Indene
OD600
Glucose (g/L)
Indene metabolites
Time (hours)
trans-indandiol (mM)
cis-indandiol (mM)
1-indenol (mM)
1-indanone (mM)
indene (mM)

Figure 4. RNA Quantitation. RNA samples harvested from *Rhodococcus* sp. I24 fermentation cultures at mid-log or early stationary phase were analyzed with an Agilent 2100 bioanalyzer to measure concentration and sample integrity. The graphic representation of peak intensity at left indicates all samples had only two major peaks (representing the ribosomal RNAs), although sample 10 underwent some sample degradation. The physiological state of each sample as well as the concentration is indicated in the table at right.



Sample number	Growth phase (OD_{600})	Aromatic substrate	Concentration (μg/uL)
1. F2 non-1	mid-log (4.9)	noninduced	0.41
2. F2 non-2	early stationary (11.5)	noninduced	0.18
3. F3 non-1	mid-log (5.1)	noninduced	0.56
4. F3 non-2	early stationary (11)	noninduced	0.17
5. F3 naph-1	mid-log (5.0)	naphthalene	0.44
6. F3 naph-2	early stationary (7.1)	naphthalene	0.34
7. F5 naph-1	mid-log (5.2)	naphthalene	0.43
8. F5 naph-2	early stationary (7.4)	naphthalene	0.35
9. F5 ind-1	mid-log (4.7)	indene	0.45
10. F5 ind-2	early stationary (10)	indene	0.21
11. F6 ind-1	mid-log (5.1)	indene	0.42
12. F6 ind-2	early stationary (9.1)	indene	0.29

Figure 5. Statistical normalization. The top left graph displays the distribution of background subtracted average pixel intensities for all spots of a single slide. Scanner normalization (top right) compresses the spread of the data by equalizing the total intensity of each fluorescent channel within the two arrays of the slide. Log transformation (bottom left) converts the data to a form that is both linear along the normal and distributed in a statistically tractable form. Lastly, expressing the intensity as a fraction of the sum of the total channel intensity per array converts the fluorescence intensity values into a measure of the physical amount of labeled cDNA bound to the microarray as a function of the total amount of cDNA bound.

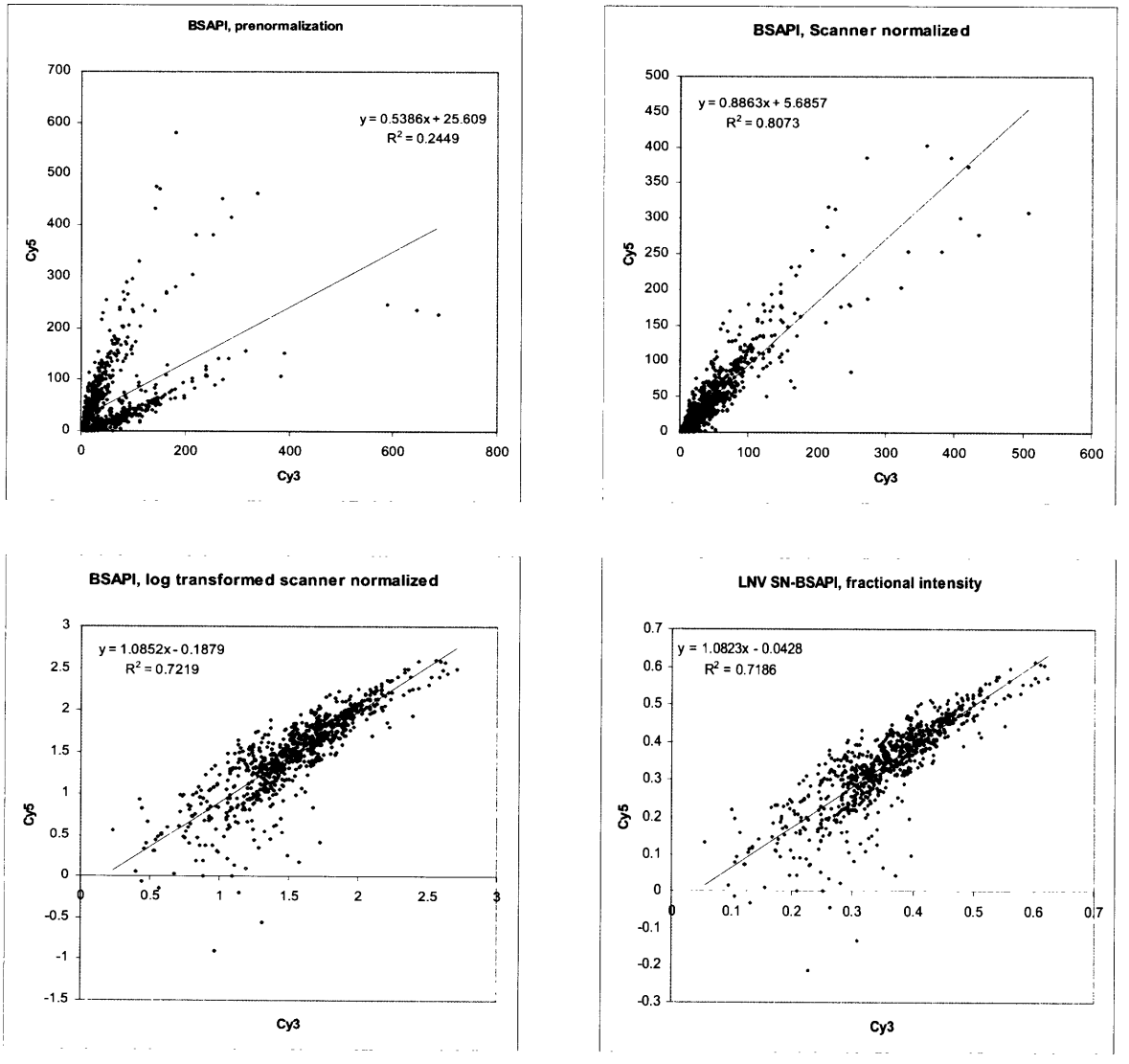


Figure 6. Average percent change in transcription of genes involved in indene bioconversion measured by trigonometric deconvolution analysis of DNA microarray data. A) noninduced vs. naphthalene induced at mid-log, B) noninduced vs. indene induced at mid-log, C) noninduced vs. naphthalene induced at early stationary, D) noninduced vs. indene induced at early stationary.

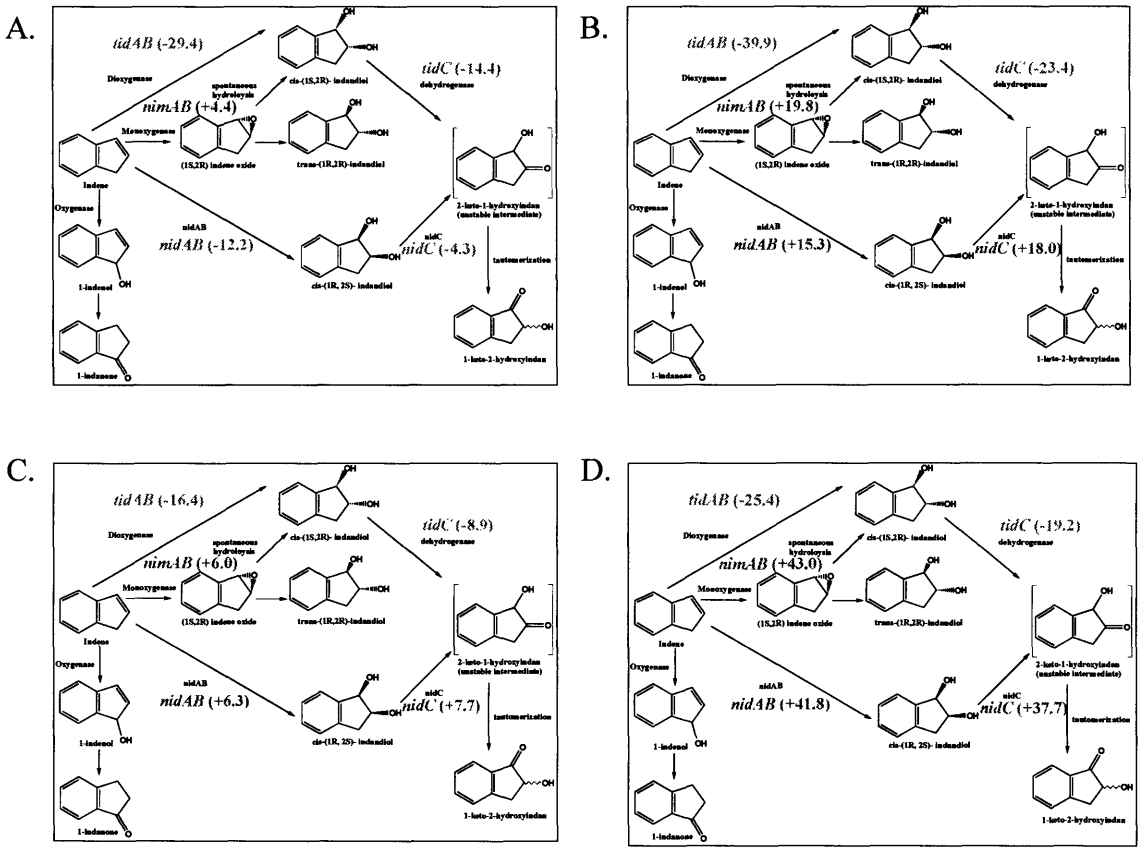


TABLE 1

Operon	ORF	Function	60mer sequence
2214	5003	Aldehyde dehydrogenase (EC 1.2.1.3)	CGGGTCGAGTCCTACATCGCCAAGGGGAAGGCCGAAGGGGCCCGGCTGACCGCCGGCGGC
2214	5004	Ferredoxin, 2Fe-2s	CGACCTGCCACGTACACGTCGACCCCGAGTACGCCGAGCTGTTCGACGCGGCCACCGACG
2214	5005	Ferredoxin--NAD(+) reductase (EC 1.18.1.3)	ACGCCCGAGCCCTGCGCAAGTGGGTCACGGAAGGCCGCACCCTCGCCATCATCGGCGGCG
2214	5204	Cytochrome P450-TERP (EC 1.14.-.-)	CCCTGTTCGCGGCCTACAAGTGGCTGCGGGAGAACAACCCGCTCGCCCAGGTACACGTCG
2214	5277	p-Cumate 2,3-dioxygenase alpha subunit (EC 1.14.12.-)	ACTGGTGGGCATCCCGGACCGCCAGGCGTACGGCGACGACCTCGACTTCGGCAAACTCGG
2214	5281	p-Cumate 2,3-dioxygenase beta subunit (EC 1.14.12.-)	TCCCGCCGGGCACACCGGGAATACCCGCACTCGGCCACCAGCCACCAGGTGTCCAACGTG
2219	2602	2-hydroxymuconic semialdehyde hydrolase	GTGCGCCGATCCCATCTCCTCACCTTCGGGCTTCCGCAATCCGGTCCCGGGCTTCCGGGC
2219	2603	Biphenyl-2,3-diol 1,2-dioxygenase (EC 1.13.11.39)	GCGAACTCGGTCCCGGTCTCGGAGGATGCCGATCTGGCGCGTGCACGGCATGTGCAGCGG
2219	3013	Ferredoxin--NAD(+) reductase (EC 1.18.1.3)	CGCGCACGATCCCCACCTCCGGGACGCGCCCCGATCTCGTGCACTACCTGCGGACGCTCG
2219	4492	short chain dehydrogenase	AGTCCTCCCGCGGCAGCATGATCTTCACCCTGTCCAACGCGGCCTTCTTCCCGGGCGGCG
2219	4493	Biphenyl dioxygenase alpha subunit (EC 1.14.12.18)	TTCGTACCGAAGGACGCGTCGCCGGAGGCCAAGGCGGGAATGCGGCGGGGCACCCAGCGG
2219	4494	Biphenyl dioxygenase beta subunit (EC 1.14.12.18)	TGGTCGGAGGAGCCCCCGTCGCGCACCCGGCGGTTGCTGACCAATGTCCGTGTCGCGCAC
2219	4495	Rieske-type ferredoxin	CCCAGCATCAGCATCTTCCTGGGAGGTATCCGACGATGACGGCTGAATCATCCAGTGAGA
2224	5992	Biphenyl dioxygenase alpha subunit (EC 1.14.12.18)	GCTGGGCCAGTACAACGAGAACAAGCGCCGCGTCGCCCGCGAGCGGGTCGGGGACCGTGC
2224	5995	Ferredoxin--NAD(+) reductase (EC 1.18.1.3)	GCGTGTTCGCTGCCGGCGACGTAGCTAACGGCCCCAACGAGTTCGCCGGTGGCCGGGTCC
2224	5997	Biphenyl-2,3-diol 1,2-dioxygenase (EC 1.13.11.39)	CGCGGATCTGGTTGTGCTCAACGACTGTGGGCACTGGGCGCCGTTCGAGCGGCCACGCGC
2224	5999	Biphenyl dioxygenase beta subunit (EC 1.14.12.18)	GCCGTCACGGTGCCGGCATCTAATCACGAATGTGCGCGTGTCGCCGTGTGCCGCGGACAG
2224	6002	Rieske-type ferredoxin	CCAGATCGAGTGTGGGTGGCACTTCGCAAAGTTCTGCATCCGGACCGGAGCAGTCACGGC
2226	3393	Cis-1,2-dihydroxycyclohexa-3,5-diene-1-carboxylate dehydrogenase (EC 1.3.1.55)	CGCGCAGAGCGAGCAGGAGAAGGGCTGGTACCAGCAGATCGTGGACCAGACCGTCGACTC
2226	4759	Benzoate 1,2-dioxygenase beta subunit (EC 1.14.12.10)	TACTACCCCAACCGCGGTGGCCTCGAGGACCGGGTGTTCCGCATCCGCACCGACCGCTCC
2226	4760	Benzoate 1,2-dioxygenase electron transfer component (EC 1.14.12.10)	CGAGCGGGAACTCGAAGCCGCCACCAGGGCAGCCGAGGAGACCGGCCTGCCGGTGTCCCT
2226	4830	Benzoate transport protein	CCTGGCACTGATCCTCCTCACCTTCGGGTTCCCGCTGCCCGTCCTGCTGACCGCGGTGGC
2226	7606	Benzoate 1,2-dioxygenase alpha subunit (EC 1.14.12.10)	TGCTGTGGATGTGGTGGGGCAACCCGCAGGACCGCCCGCTCTTCCCCCGCAAGGACGAGC
2247	5468	4-hydroxyphenylacetate 3-monooxygenase (EC 1.14.13.3)	GGCGGACTGCTCTACCAGCCGGCCGACGTCAGTGCGTTCGACTCCCCCATCGCGTCGGAC
2247	5469	Catechol 2,3-dioxygenase (EC 1.13.11.2)	CTGAGGGGCCGCAGTACATTCATCCGGCGCGAGCCGCCCATCGCGGCGGGATCCGGGCAC
2247	5471		GAATCCGTGCACTGCGTCGTTGACGGCAGCGGCTACTCCGAGATCGGCGGGAAGACCCTG
Nid	6043	naphthalene dioxygenase large subunit	CCGCGCGGAGATGGGGAACGCCTCACACTTCCGGTGCCCCTACCACGGCTGGACCTACAG
Nid	6044	Trans-O-hydroxybenzylidenepyruvate hydratase-aldolase (EC 4.2.1.-)	CCCCCATTTCTGACAGCACCCGAGGCCTATGCCGAGGGGGCGCGCGAAAACGGACGTCGG
Nid	6047	cis-naphthalene dihydrodiol dehydrogenase	TATGACGCGGATGGCGGGTCTCGCGCAGTGAGGGTCGGCGCGTGGGTGGGCCACTCCTCC
Nid	6049	naphthalene dioxygenase small subunit	CGTCGCGCTCACGGCACTTCGTCACCAACGTTCAAGTCGCACCGGGCGATAGCGAGGACG
Nid	6057	NidC diol dehydrogenase	CTGCTTGATCAGACCACGATAATGACACACAACCTCGCCCTGATCATGTGACGGAGGAGC
Nim	617	Styrene monooxygenase large component (EC 1.14.13.-)	AGGTCATCGAGGCCGATCCGTACTCGTGGCTGCGCGGGGCCGTCACCCCGACGGTGCGGC
Nim	4636	NimR protein	GCGTCCCCGCCCCTGAGTTGAATCCGGCCGAATGGCGCGTGCCCGTCAGGGCGAACACTG
Nim	5714	2,3-dihydroxyphenylpropionate 1,2-dioxygenase (EC 1.13.11.-)	GATCGGTCCGCCAGCAGCGGGTTATCGACACCGCCAAGGCCTTCACCAGGGGTGAGGCCG
Nim	5715	acetaldehyde dehydrogenase	CTTGCGGGGCCCGGGCCTCACCATCGAGGATGCTGCAGCCGCAGTGGGGCAGGTGCTGCG
Nim	5717	4-hydroxy-2-oxovalerate aldolase (EC 4.1.3.-)	CTCGGGTGGACAGCTGGCCATGAACGACGTACGAGCGCGGATGCGCGCCTACCGAGGGGT
Nim	5719	phenol 2-hydroxylase component B	GGTGCCGACCAGGAAACCGTTAGCCGGCACTTCAGTGGGCGCCCCGACCTCCTTCCAGAT
Tid	6111	Biphenyl dioxygenase alpha subunit (EC 1.14.12.18)	CAGGCCTTCCCCGGGCTGAGGAAAGAAGATTGGGGCCCGCTACAGGCTCGCGTCGAGACC
Tid	6112	Biphenyl dioxygenase system ferredoxin--NAD(+) reductase component (EC 1.18.1.3	CGCGAGAGTTCGCCATGGCGACCCGCCTTGTCGAGAGCGGTGCTCAGGTGGGCCGGGACG
Tid	6113	Biphenyl-2,3-diol 1,2-dioxygenase (EC 1.13.11.39)	GTCGGCTCGGGACGCCGGGTGGCCAATCTCCTTCGGCCATCACACACGGCAACAGGAAAG
Tid	6114	BIPHENYL-2,3-DIHYDRO-2,3-DIOL DEHYDROGENASE (EC 1.3.1.-)	GCGCGATTATCAACTGTGACGGCGGGATGGGGGTGCGCGGCTTGGCCGAGACGGCCGGCG
Tid	6115	Biphenyl dioxygenase beta subunit (EC 1.14.12.18)	AATACGCACATTTCGACGACAATGCGCAGATGATGCGAGGGCGCCTGCGCAAGATCACTT
Tid	6120	Biphenyl dioxygenase system ferredoxin component	CGCCAACAACCTCAGCTTCTTCTTCTAGGAGTGACGATGGCCCTCACAAAGATATGCAGC

Table 1. Probe sequences and descriptions. Open reading frames are grouped by operons as they are organized in the genome (http://ergo.integratedgenomics.com/ERGO/) and coded by the Integrated Genomics number designation. Putative functions describe either the Integrated Genomics functional annotation assignment or the predicted function based on automated BLASTx sequence alignment analysis with the CAPASA functional annotation application (Chapter 2).

References

Bell, KS, JC Philp, DW Aw and N Christofi (1998). The genus *Rhodococcus*. J Appl Microbiol **85**(2): 195-210.

Bozdech, Z, J Zhu, MP Joachimiak, FE Cohen, B Pulliam and JL DeRisi (2003). Expression profiling of the schizont and trophozoite stages of *Plasmodium falciparum* with a long-oligonucleotide microarray. Genome Biol **4**(2): R9.

Brody, JP, BA Williams, BJ Wold and SR Quake (2002). Significance and statistical errors in the analysis of DNA microarray data. Proc Natl Acad Sci U S A **99**(20): 12975-8.

Brown, CS, PC Goodwin and PK Sorger (2001). Image metrics in the statistical analysis of DNA microarray data. Proc Natl Acad Sci U S A **98**(16): 8944-9.

Chartrain, M, Jackey, B., Taylor, C., Sandford, V., Gbewonyo, K., Lister, L., Dimichele, L., Hirsch, C., Heimbuch, B., Maxwell, C., Pascoe, D., Buckland, B., Greasham, R. (1998). Bioconversion of indene to *cis* (1S,2R) indandiol and *trans* (1R,2R) indandiol by *Rhodococcus* species. Journal of Fermentation and Bioengineering **86**(6): 550-558.

Datta, S (2003). Comparisons and validation of statistical clustering techniques for microarray gene expression data. Bioinformatics **19**(4): 459-66.

Dudley, AM, J Aach, MA Steffen and GM Church (2002). Measuring absolute expression with microarrays with a calibrated reference sample and an extended signal intensity range. Proc Natl Acad Sci U S A **99**(11): 7554-9.

Finnerty, WR (1992). The biology and genetics of the genus *Rhodococcus*. Annu Rev Microbiol **46**: 193-218.

Goodfellow, M, G Alderson and J Chun (1998). Rhodococcal systematics: problems and developments. Antonie Van Leeuwenhoek **74**(1-3): 3-20.

Guillouet, S, AA Rodal, G An, PA Lessard and AJ Sinskey (1999). Expression of the *Escherichia coli* catabolic threonine dehydratase in *Corynebacterium glutamicum* and its effect on isoleucine production. Appl Environ Microbiol **65**(7): 3100-7.

Hughes, DL, GB Smith, J Liu, GC Dezeny, CH Senanayake, RD Larsen, TR Verhoeven and PJ Reider (1997). Mechanistic Study of the Jacobsen Asymmetric Epoxidation of Indene. J Org Chem **62**(7): 2222-2229.

Larkin, MJ, R De Mot, LA Kulakov and I Nagy (1998). Applied aspects of *Rhodococcus* genetics. Antonie Van Leeuwenhoek **74**(1-3): 133-53.

Loos, A, C Glanemann, LB Willis, XM O'Brien, PA Lessard, R Gerstmeir, S Guillouet and AJ Sinskey (2001). Development and validation of *Corynebacterium* DNA microarrays. Appl Environ Microbiol **67**(5): 2310-8.

O'Brien, XM, JA Parker, PA Lessard and AJ Sinskey (2002). Engineering an indene bioconversion process for the production of cis-aminoindanol: a model system for the production of chiral synthons. Appl Microbiol Biotechnol **59**(4-5): 389-99.

Orru, RV, A Archelas, R Furstoss and K Faber (1999). Epoxide hydrolases and their synthetic applications. Adv Biochem Eng Biotechnol **63**: 145-67.

Pan, W (2002). A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. Bioinformatics **18**(4): 546-54.

Park, T, SG Yi, S Lee, SY Lee, DH Yoo, JI Ahn and YS Lee (2003). Statistical tests for identifying differentially expressed genes in time-course microarray experiments. Bioinformatics **19**(6): 694-703.

Quackenbush, J (2002). Microarray data normalization and transformation. Nat Genet **32 Suppl**: 496-501.

Reider, PJ (1997). Advances in AIDS chemotherapy: The asymmetric synthesis of CRIXIVAN(R). Chimia **51**(6): 306-308.

Senanayake, CH, GB Smith, KM Ryan, LE Fredenburgh, J Liu, FE Roberts, DL Hughes, RD Larsen, TR Verhoeven and PJ Reider (1996). The role of 4-(3-phenylpropyl)pyridine N-oxide (P3NO) in the manganese-salen-catalyzed asymmetric epoxidation of indene. Tetrahedron Letters **37**(19): 3271-3274.

Sokal, RR, Rohlf, F.J. (1987). Introduction to Biostatistics. New York, New York, W.H. Freeman and Company.

Stafford, DE, KS Yanagimachi, PA Lessard, SK Rijhwani, AJ Sinskey and G Stephanopoulos (2002). Optimizing bioconversion pathways through systems analysis and metabolic engineering. Proc Natl Acad Sci U S A **99**(4): 1801-6.

Swanson, PE (1999). Dehalogenases applied to industrial-scale biocatalysis. Curr Opin Biotechnol **10**(4): 365-9.

Taylor, JR (1997). An Introduction to Error Analysis. Sausalito, California, University Science Books.

Treadway, SL, KS Yanagimachi, E Lankenau, PA Lessard, G Stephanopoulos and AJ Sinskey (1999). Isolation and characterization of indene bioconversion genes from *Rhodococcus* strain I24. Appl Microbiol Biotechnol **51**(6): 786-93.

Vacca, JP, BD Dorsey, WA Schleif, RB Levin, SL McDaniel, PL Darke, J Zugay, JC Quintero, OM Blahy, E Roth and et al. (1994). L-735,524: an orally bioavailable human immunodeficiency virus type 1 protease inhibitor. Proc Natl Acad Sci U S A **91**(9): 4096-100.

Warhurst, AM and CA Fewson (1994). Biotransformations catalyzed by the genus Rhodococcus. Crit Rev Biotechnol **14**(1): 29-73.

Yamada, H and M Kobayashi (1996). Nitrile hydratase and its application to industrial production of acrylamide. Biosci Biotechnol Biochem **60**(9): 1391-400.

Yanagimachi, KS, DE Stafford, AF Dexter, AJ Sinskey, S Drew and G Stephanopoulos (2001). Application of radiolabeled tracers to biocatalytic flux analysis. Eur J Biochem **268**(18): 4950-60.

Chapter IV

Future Perspectives

Rhodococcus sp. I24 as a Manufacturing Platform

The genus *Rhodococcus* is best known for its range of metabolic activities (Finnerty 1992). A very brief list includes production of both flocculents and emulsifiers, biotransformation of short and long chain aliphatic hydrocarbons, complete metabolism of polycyclic aromatic hydrocarbons to CO_2, degradation of halogenated hydrocarbons including polychlorinated biphenyls (PCBs), and transformation of nitriles (Warhurst et al. 1994). The diversity of metabolic activity within this class of bacteria has made it the focus of exploration for novel sources of biocatalytic reagents by many groups.

Rhodococcus sp. I24 was isolated by researchers at Merck and Co. (Rahway, NJ) through an enrichment selection for organisms capable of consuming naphthalene or toluene as a sole carbon source (Chartrain et al. 1998). The priority of these researchers was to develop a strain for the manufacture of 2R-indandiol for the production of the HIV-1 protease inhibitor Crixivan™. Research is ongoing in our research group to achieve this goal, but other exploration is ongoing to fully prospect the value of this strain. Cloning studies revealed the presence of several polycyclic aromatic hydrocarbon dioxygenases responsible for the naphthalene and toluene metabolizing activity that first gained notice (Chartrain et al. 1998; Treadway et al. 1999); O'Brien, unpublished results), but questions remained about what was still unknown. Initial answers arrived in 2000 as a draft version of the *Rhodococcus* sp. I24 genome determined by Integrated Genomics, Inc. (Chicago, IL). One aspect of the research described in this thesis was the development and validation of an application for automated functional annotation transfer, the Consensus Annotation by Phylogeny Anchored Sequence Alignment program (CAPASA). The combined information derived from the Integrated Genomics

ERGO™ database (Overbeek et al. 2003) and CAPASA, along with genetic tools, revealed a vast array of synthetic potential within this one strain of *Rhodococcus*.

The genome sequence and functional annotation of the *Rhodococcus* sp. I24 genome revealed many standard activities that would be expected in any bacterium. Table 1 lists the number of representative open reading frames determined by Integrated genomics from a number of classes.

104	non-ribosomal peptide synthetase
34	cytochrome P450
22	PAH dioxygenase small subunit
16	PAH dioxygenase large subunit
10	aromatic extradiol (ring cleaving) dioxygenase
5	polyhydroxyalkanoate (PHA) polymerase
2	polyhydroxyalkanoate (PHA) depolymerase

Table 1

The multitude of non-ribosomal peptide synthetase subunits is a potential source of novel polyketides and peptide based antibiotics (Harris et al. 1974; Shen 2003), while the polycyclic aromatic hydrocarbon (PAH) dioxygenases could produce a variety of chiral synthons for chemical and pharmaceutical manufacturing (O'Brien et al. 2002). Lastly, multiple polyhydroxy-alkanoate polymerase and depolymerase enzymes could serve as novel reagents for the production of biodegradable plastics with a range of properties desirable in many materials (Madison et al. 1999).

The next steps in the development of *Rhodococcus* sp. I24 include completion of genome sequencing, annotation of DNA sequences, and functional annotation assignment (in part with CAPASA). Once a high content draft of the complete genome is available, full genome microarrays can be designed for high-resolution time dependent

measurement of gene transcription during indene, naphthalene, and toluene metabolism using trigonometric deconvolution analysis of microarray data. Repeating the fermentation cultures described earlier with sampling for aromatic hydrocarbon and RNA on the order of 15-30 minutes will allow the fine resolution of transcription at the onset of each of the indene metabolite activity inductions, as well as a better correlation of global gene activity with aromatic metabolism. Such an experiment will also highlight the value of trigonometric deconvolution analysis to maximize measurement flexibility across any set of physiological conditions with a minimal number of hybridizations, which is often a factor when designing exceptionally high priced microarray experiments. Full genome array analysis will also enable the identification of genes involved in other metabolic processes of interest, allowing further analysis by cloning, gene knock outs, and over-expression. Genome scale sequencing and computational determination of gene function will open the door for the complete exploration of *Rhodococcus* sp. I24, an organism that was initially intractable to standard methods of genetic analysis and manipulation.

The Future of Biocatalysis

The interplay of multiple research disciplines will continue to be the driving force in the development of biocatalysis processes for the synthesis of novel products, in contrast to biotechnology where the biological material *is* the product. Biological systems of purified enzymes or whole cells will be rationally engineered to achieve a specific goal, with predictable results. Advances in many areas must continue to be made to achieve the full potential of this technology.

Information Transfer

The first step in achieving the challenge of biocatalysis by design will be the expansion and refinement of knowledge systems for biological reactions. Databases like the University of Minnesota Biocatalysis/ Biodegradation Database (Ellis et al. 2003) and the Kyoto Encyclopedia of Genes and Genomes (KEGG; (Kanehisa et al. 2002) are the likely foundations for international libraries of enzyme catalyzed reactions. These virtual knowledge stores must expand their holdings to include the maximum amount of information about a biological reagent as possible. Information about the full substrate utilization range will be needed, in addition to the normal physiological function. High throughput fluidic analysis systems will be needed to measure the reaction kinetics of native and modified forms of the enzyme in a range of solvents and temperatures. Lastly, intramolecular regulatory mechanisms, such as feedback inhibition, must be identified and characterized as completely as possible. As more information becomes available such a resource can be used for the virtual design of a chemical synthetic process for construction and implementation on the laboratory benchtop, pilot plant, or factory.

Reagent Availability

Chemical catalysts can be ordered from any of a number of vendors. The availability of biocatalysts must be just as widespread for the field to achieve its full potential. Physical libraries of enzyme DNA cassettes must mirror the virtual libraries of enzyme reactions stored in computer databases. The American Type Culture Collection (Manassas, VA) serves as a good model of how such a collection could be organized and maintained. As new enzymes are isolated, engineered, and characterized they need to be deposited in a centralized facility for easy access by others. Such an agency could also

serve as a clearinghouse for distribution of compensation to the depositor to encourage the continued sharing of resources. Biological process design should approach the ease of reverse engineering a reaction scheme from a desired product to ordering the necessary enzyme components in easy to assemble standardized DNA shuttle vectors.

Training

Researchers and engineers who are equally comfortable designing a petrochemical refinery platform and calculating the specific growth rate of cultures metabolizing crude oil will lead the future of biotechnology and biocatalysis. Cross-disciplinary collaboration and training should be the norm instead of the exception. Early adoption could include the integrating introductory level courses in "external" fields as electives during graduate academic education. Expansion and integration of computer science, engineering, and biology will occur as necessity requires. Ultimately, depth of knowledge and tightly focused expertise will be balanced with creative problem solving capabilities and an ability to integrate new information.

Conclusions

The continuing evolution of biology will incorporate aspects of computer science, chemical engineering, and mathematics in innovative ways to allow the design and development of synthetic tools for biocatalysis. Advances in affordable computation have enabled the management, manipulation, and analysis of the vast amounts of data generated by global analysis tools like DNA microarrays, full genome sequence comparison, and high throughput robotic assays. Biology is a science of complex systems, possibly the most complex in existence. Biologists of the future will study these systems in their native complex state with analytical tools constructed from multiple

fields. The synthesis of knowledge created by the fusion of such skills will lead to the ultimate goal of rational design of biological processes for the manufacture of value added compounds, materials, and therapeutics.

References

Chartrain, M, Jackey, B., Taylor, C., Sandford, V., Gbewonyo, K., Lister, L., Dimichele, L., Hirsch, C., Heimbuch, B., Maxwell, C., Pascoe, D., Buckland, B., Greasham, R. (1998). Bioconversion of indene to *cis* (1S,2R) indandiol and *trans* (1R,2R) indandiol by *Rhodococcus* species. Journal of Fermentation and Bioengineering **86**(6): 550-558.

Ellis, LB, BK Hou, W Kang and LP Wackett (2003). The University of Minnesota Biocatalysis/Biodegradation Database: post-genomic data mining. Nucleic Acids Res **31**(1): 262-5.

Finnerty, WR (1992). The biology and genetics of the genus Rhodococcus. Annu Rev Microbiol **46**: 193-218.

Harris, TM, CM Harris and KB Hindley (1974). Biogenetic-type syntheses of polyketide metabolites. Fortschr Chem Org Naturst **31**(0): 217-82.

Kanehisa, M, S Goto, S Kawashima and A Nakaya (2002). The KEGG databases at GenomeNet. Nucleic Acids Res **30**(1): 42-6.

Madison, LL and GW Huisman (1999). Metabolic engineering of poly(3-hydroxyalkanoates): from DNA to plastic. Microbiol Mol Biol Rev **63**(1): 21-53.

O'Brien, XM, JA Parker, PA Lessard and AJ Sinskey (2002). Engineering an indene bioconversion process for the production of cis-aminoindanol: a model system for the production of chiral synthons. Appl Microbiol Biotechnol **59**(4-5): 389-99.

Overbeek, R, N Larsen, T Walunas, M D'Souza, G Pusch, E Selkov, Jr., K Liolios, V Joukov, D Kaznadzey, I Anderson, et al. (2003). The ERGO genome analysis and discovery system. Nucleic Acids Res **31**(1): 164-71.

Shen, B (2003). Polyketide biosynthesis beyond the type I, II and III polyketide synthase paradigms. Curr Opin Chem Biol **7**(2): 285-95.

Treadway, SL, KS Yanagimachi, E Lankenau, PA Lessard, G Stephanopoulos and AJ Sinskey (1999). Isolation and characterization of indene bioconversion genes from Rhodococcus strain I24. Appl Microbiol Biotechnol **51**(6): 786-93.

Warhurst, AM and CA Fewson (1994). Biotransformations catalyzed by the genus Rhodococcus. Crit Rev Biotechnol **14**(1): 29-73.

Chapter V.

Acknowledgments

I would like to thank many people for the support, assistance, patience, and love given to me throughout my graduate career at the Massachusetts Institute of Technology. It has been a long time and many of my colleagues have come and gone, but they will never leave my heart even if they have left my mind.

I would like to say thank you to my family for always being there for me, inspiring me to carry on to become the first to complete graduate school. I can only hope to serve as an inspiration to the future generations of nieces, nephews, and cousins to always pursue their dreams in spite of circumstances.

I would like to thank my undergraduate advisor Dr. Lynnette Padmore for never letting me get by with anything less than 100%, even when less would have gotten an A.

I would like to thank my graduate advisor Prof. Anthony Sinskey for having faith in me, even when my path wasn't clear to him. Thank you for allowing me to pursue my own interests, and in the process re-find my own love of science.

Thank you to members of the Sinskey lab (past and present) including all the graduate students, post-docs, technicians, diploma students, UROPs, MSRP students, a few over-achieving high schoolers, and one grand daughter. If it weren't for all of you I would have given up on this whole thing a long time ago.

In particular I would like to give a special thanks to the following members of the Sinskey lab for contributions above and beyond the call of professional academic relationships. Dr. Philip A. Lessard, the best educator at MIT, loving husband of Jennifer, father of Joe, Steven, and Kate, and a model human being. Xian O'Brien, one of the smartest people I know, and someone who just makes the world better. Good luck in graduate school at Brown University. Dr. Laura Willis, for making sense of

everything, even when nothing made sense. Amie J. Strong, for listening while I had to think out loud. Zofia Gajdos and Binbin Wang, for being so patient learning from me while I learned to teach you. Jennie Cho, for just being Jennie-son.

I would also like to thank: Nathan, Melina, Diana, Adriane, Ellen, Aretha, Devin, Annette, Whei, Caitlin, Nathalie, Allison, Molly, Diviya, Irene, Elaine, Amanda, Tennyson, Andrea, Geeta, Lorien, Nancy, Horst, Supriya, Annet, Josh, Sushil, Vera, Jina, Dan, Sheri, Sladjana, Sheila, Binbin, Kurt, Chong Yi, Kevin, Joe, Adam, Jessica, Alina, Kazuhiko, Vu, Paolo, Joe, Robert, Erich, Eudean, and whoever else has worked in lab while I've been there but can't remember right now for making the entire lab environment more bearable.

A special thank you to the members of my thesis committee: Profs. Robert Sauer, Alan Grossman, and Graham Walker of the Massachusetts Institute of Technology, and Prof. John Archer or Cambridge University. Your insights, advice, questions, and criticisms have contributed to making me look at the world in new ways.

Last, but definitely not least, I would like to thank the National Science Foundation, The Bioprocess Engineering Center, The MIT Provost's Office, and E.I. DuPont de Nemours and Company for paying me $151,375.00 over all these years.

I H T F P

1996 - 2004

JAP