Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Molecular Catalysis B: Enzymatic 31 (2004) 87–96
JOURNAL OF MOLECULAR CATALYSIS B: ENZYMATIC
www.elsevier.com/locate/molcatb

A comparative molecular field analysis of the biotransformation of sulfides by *Rhodococcus erythropolis*

Jarrod B. French*, Giles Holland, Herbert L. Holland, Heather L. Gordon*

Department of Chemistry and Centre for Biotechnology, Brock University, St. Catharines, ON, Canada, L2S 3A1

Received 12 August 2004; accepted 12 August 2004
Available online 16 September 2004

Abstract

A comparative molecular field analysis (CoMFA) was used to model the efficacy with which the *Rhodococcus erythropolis* mono-oxygenase, DszC, catalyzes the enantioselective sulfoxidation of a broad range of substrates. Experimentally determined values of both the yield and enantiomeric excess for this reaction were employed to create these CoMFA models. A highly predictive CoMFA model was constructed for the prediction of enantiomeric excess of the sulfoxide product. The predictive ability of the model was demonstrated by both cross-validation of the training set ($q^2 = 0.74$) and for an external test set of substrates. The enantiomeric excesses of the members of the test set, which also included two amino acid sulfides that were structurally distinct from the membership of the training set, were predicted well by the CoMFA model. Product yield was not modelled well by any CoMFA model. Different models comparing the likely bioactive conformations of the substrates suggest that most compounds assume an 'extended' conformation upon binding. Contour diagrams illustrating significant substrate–enzyme interactions suggest that the model, which predicts the enantiomeric excess, is consistent with previous conclusions regarding the effect of various substrate substitutions on the enantiopurity of the product of the biotransformation.

Keywords: *Rhodococcus erythropolis*; CoMFA; 3D-QSAR; Biocatalysis; Sulfoxide

1. Introduction

Eubacteria of the genus *Rhodococcus* are a diverse group of microorganisms exhibiting broad metabolic diversity that are commonly found in many environmental niches from soils to seawater and as plant and animal pathogens [1–3]. The economic importance of this bacterium is becoming increasingly apparent as knowledge of its genetics and biochemistry accumulates. One current area of research of *Rhodococcus* strains is the biocatalytic desulfurization of fossil fuels [2–4]. This process employs the bacterium to remove the sulfur from petroleum products without degrading the fuel value.

The desulfurization pathway of *Rhodococcus erythropolis* IGTS8 has been extensively studied and its enzymes characterized [5–7]. This pathway is illustrated in Fig. 1 for the desulfurization of dibenzothiophene, one of the major components of middle-distillate petroleum [3,8]. This four-step process involves three enzymes, which are found in the plasmid-encoded dibenzothiophene desulfurization operon. Of the three enzymes involved, two are cytoplasmic mono-oxygenases (DszA and DszC), while the third (DszB) is a desulfinase.

Several mutant strains of this bacterium have been engineered in an attempt to optimize its ability to remove sulfur [4,8]. One particular mutant, BKO-53, engineered by Energy Biosystems Corporation [7,8] expresses only the mono-oxygenase DszC. This mutant has been recently examined for its ability to function as a stereospecific biocatalyst for the oxidation of sulfides [9,10]. Results have shown that this strain of *Rhodococcus* is able to convert sulfides to the corresponding sulfoxides in moderate to high yields with good stereoselectivity [9,10].

Chiral sulfoxides are of interest for a number of reasons. Sulfoxide derivatives of naturally occurring amino acids can

* Corresponding authors. Tel.: +1 905 688 5550x4604; fax: +1 905 682 9020.
E-mail addresses: jf00aa@brocku.ca (J.B. French), gordonh@brocku.ca (H.L. Gordon).

1381-1177/$ – see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.molcatb.2004.08.001

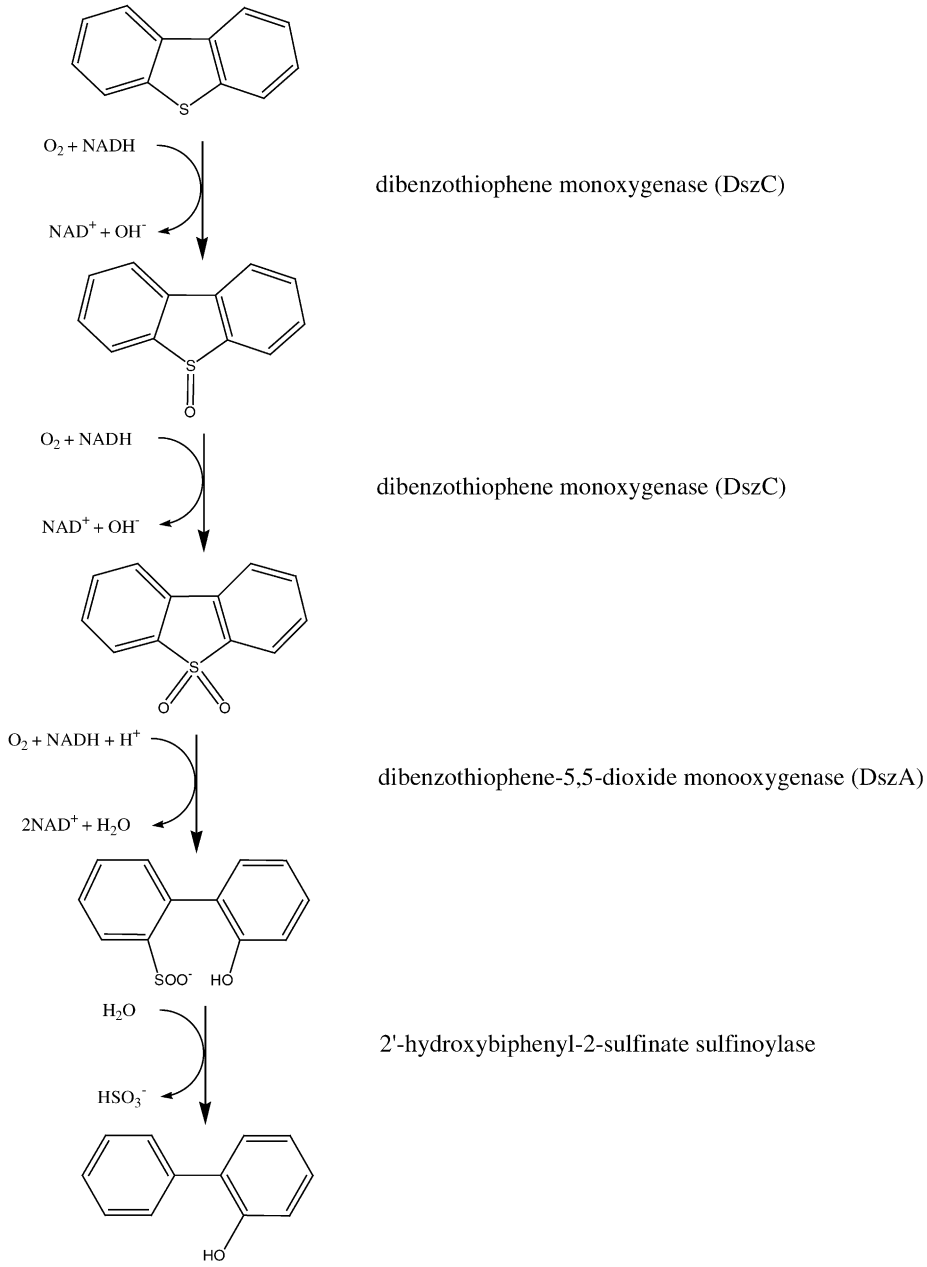

Fig. 1. The metabolic pathway for the desulfurization of DBT to HBP and sulfite. DszC catalyzes the conversion of dibenzothiophene (DBT) to the sulfoxide dibenzothiophene 5-oxide (DBTO) and further to the sulfone dibenzothiophene 5,5-dioxide ($DBTO_2$). DszA then further degrades $DBTO_2$ to 2-(2′-hydroxyphenyl)benzene sulfinate (HBPSi). Finally, DszB catalyzes the conversion of HBPSi to 2-hydroxybiphenyl (HBP) and sulfite (SO_3^{2-}) [8].

act to regulate cholesterol catabolism [11] and have antibiotic properties [12]. Chiral sulfoxides are valuable as asymmetric starting materials [13–16], and as chiral auxiliaries [15] in organic synthesis. They have been chemically synthesized using such reagents as oxaziridines or peroxide–metal complexes [17], and more recently, using chiral binaphthol [18] or diphenylethan-1,2-diol [19]. Isolated enzymes such as peroxidases, haloperoxidases [16,20,21] and mono/di-oxygenases [16,21,22] can also be employed for the stereospecific oxidation of sulfides. Whole cell biocatalysts, including bacteria such as *Rhodococcus*, *Pseudomonas putida*, and genetically modified *E. coli* [21], or fungi such as *Aspergillus* sp., *Helminthosporium* sp., and *Mortierella isabellina* [21,23] can also be used to generate chiral sulfoxides.

In a previous article published in this journal, the yields and enantiomeric excesses (e.e.) for the biocatalytic sulfur oxidation of a series of substrates by *R. erythropolis* IGTS8 were reported [9]. This paper extends the work reported therein by creating a model of the active site of the enzyme responsible for the biocatalytic sulfur oxidation. The goal of this work was to construct a three-dimensional structure-activity relationship (3D-QSAR) of the sulfur oxidizing enzyme DszC of *R.*

erythropolis IGTS8. While both two- and three-dimensional active site models have been constructed for a similar enzyme in the fungal biocatalyst *Helminthosporium* sp. NRRL 4671, such models can only provide qualitative measures of substrate diversity [21,24]. Comparative molecular field analysis (CoMFA) [25–28], a method that can be used to construct a 3D-QSAR, uses intermolecular potential energies between an atomic probe and a series of substrates to model interactions within the binding site of an enzyme. In this paper we employ a CoMFA to create both a predictive model of e.e. of the oxidation reaction of DszC and a qualitative representation of the important interactions that may exist within that enzyme's active site.

2. Results and discussion

In order to create a CoMFA model to predict the outcome of the sulfoxidation reaction by *R. erythropolis*, an appropriate training set of known substrates was established. The set of 26 compounds used to construct the final CoMFA model is listed in the top portion of Table 1. The e.e. values predicted by the model are listed alongside the experimental values determined by Holland et al. [9]. The good agreement between the experimental and predicted values of e.e. is an initial indication of the strong predictive ability of the model. It should be noted that in this and all following tables, the e.e. values correspond to the fractional excess of the R enantiomer.

Also shown in Table 1 are the data for four additional compounds (38, 51, 41 and 45) originally included in the training set that were later removed as outliers. Two of these compounds were the only two members of the training set that were found to produce predominantly the S enantiomer in the desulfurization reaction (compounds 38 and 51). The other two compounds were the only two compounds of the training set that were dibenzyl sulfides (41 and 45). The CoMFA model developed using the 26 member training set is unable to predict the e.e. of the product of the enzymatic sulfoxidation for these four compounds, as indicated by their large residuals in Table 1. In order to produce a successful prediction of e.e. for compounds 38, 51, 41, and 45, they will have to be treated in a different fashion than the 26 members of the training set, either by aligning them in a different orientation, using a different conformation, or by developing a unique CoMFA model.

Table 1
Training set for final model

Compound number[a]	Formula	Experimental e.e.[a]	Predicted e.e.	Residual[b]
1	$PhSCH_3$	0.02	0.01	−0.01
3	*m*-CH_3PhSCH_3	0.10	0.10	0.00
4	*o*-CH_3PhSCH_3	0.04	0.04	0.00
5	*p*-$CH_3OPhSCH_3$	0.22	0.14	−0.08
6	*p*-$FPhSCH_3$	0.63	0.63	0.00
7	*p*-$ClPhSCH_3$	0.72	0.73	0.01
9	*p*-NO_2PhSCH_3	0.99	0.98	−0.01
10	*p*-$CNPhSCH_3$	0.85	0.85	0.00
11	1-NaphthylSCH_3	0.25	0.31	0.06
12	2-NaphthylSCH_3	0.62	0.56	−0.06
13	$PhCH_2SCH_3$	0.26	0.33	0.07
19	*p*-,*i*-$C_3H_7PhCH_2SCH_3$	0.49	0.51	0.02
21	*p*-$CH_3OPhCH_2SCH_3$	0.27	0.29	0.02
25	*p*-$CH_3COPhCH_2SCH_3$	0.22	0.21	−0.01
27	*p*-$FPhCH_2SCH_3$	0.62	0.64	0.02
28	*p*-$CF_3PhCH_2SCH_3$	0.85	0.88	0.03
29	*p*-$ClPhCH_2SCH_3$	0.65	0.72	0.07
30	*p*-$BrPhCH_2SCH_3$	0.76	0.73	−0.03
31	*p*-$NO_2PhCH_2SCH_3$	0.76	0.71	−0.05
32	*m*-$NO_2PhCH_2SCH_3$	0.52	0.56	0.04
33	*o*-$NO_2PhCH_2SCH_3$	0.70	0.59	−0.11
40	*p*-$BrPhSCH_2CN$	0.94	0.95	0.01
47	2-PyridylCH_2SCH_3	0.24	0.25	0.01
48	4-PyridylCH_2SCH_3	0.54	0.55	0.01
54	2-ThiopheneCH_2SCH_3	0.11	0.07	−0.04
56	3-ThiopheneCH_2SPh	0.34	0.35	0.01
38[c]	$PhSCH_2CN$	−0.13[d]	0.27	0.40
51[c]	2-PyridylCH_2SPh	−0.45[d]	0.40	0.85
41[c]	$PhCH_2SPh$	0.99	0.16	−0.83
45[c]	*p*-CH_3PhCH_2SPh	0.82	0.12	−0.70

[a] Compound numbers and e.e. values correspond to those reported in Tables 1, 2 and 4 in Holland et al. [9].
[b] Residual = predicted e.e. − experimental e.e.
[c] Compounds determined to be outliers that were removed from the training set.
[d] Compounds 38 and 51 were measured to have an e.e. of the S enantiomer and are thus reported as negative values relative to the others.

CoMFA is a shape-dependent technique [25] and therefore calculated field values are highly dependent upon the orientation and conformation of the substrates under consideration. Additionally, it is widely believed that ligand–protein interactions usually occur when the ligand is at or near one of its local minimum energy conformations [25,26]. Following these assumptions, all the members of the training set were individually aligned to each of the two low energy conformations of the template molecule ($PhSCH_2Ph$: compound 41, Table 3 in [9]). This template was chosen because of its high activity and because it possessed only two minimum energy conformations, which simplified the alignment of the members of the training set. It was disappointing therefore, that the e.e. was not well predicted for the chosen template (see Table 1). However, this does not detract from the success of the ensuing model, since the template merely provides a common geometric framework for substrate alignment and does not otherwise influence the predictive model if removed from the training set. The low energy conformations of the template are shown in Fig. 2, which shows the molecule in both the extended and folded conformations. The structural variation encompassed by the 26 members of the training set

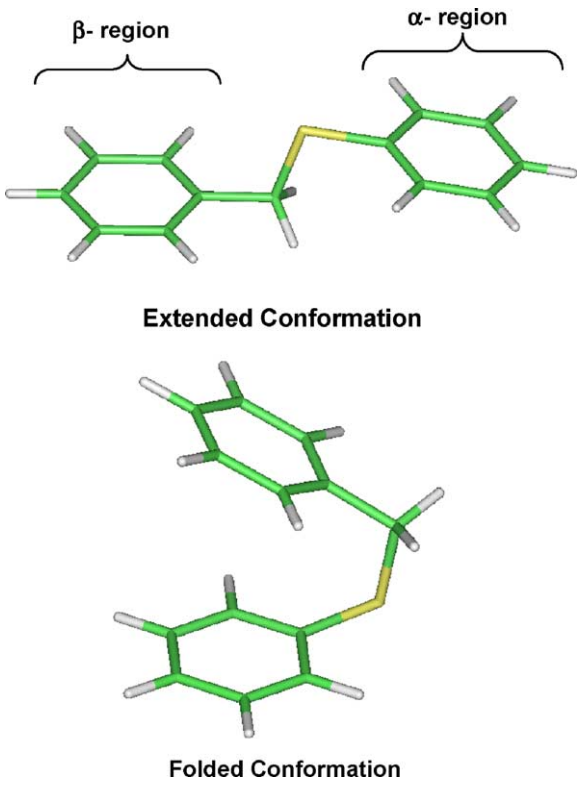

Fig. 2. Representation of the extended and folded conformations of the template molecule, $PhSCH_2Ph$. For ease of discussion within the text, the aromatic ring bonded in a position alpha to the sulfur is denoted as the α-region and the aromatic ring bonded in a position beta to the sulfur is denoted the β-region. Extended conformation dihedral angles: χ_1 ($C_{\alpha 2} - C_{\alpha 1} - S - C$) = 90.0°, χ_2 ($C_{\alpha 1} - S - C - C_{\beta 1}$) = 180.0°, χ_3 ($S - C - C_{\beta 1} - C_{\beta 2}$) = 90.1°; folded conformation dihedral angles: χ_1 = 82.8°, χ_2 = 50.4°, χ_3 = 78.0°. Figure prepared using InsightII [30].

Table 2
Optimized CoMFA parameter set

Conformation	Extended
Activity	e.e.
Grid dimensions[a] (Å)	±10
Step size[b] (Å)	2
Energy cutoff[c] (kcal/mol)	20
σ^2 – variance cutoff[d] (kcal/mol)^2	1
Number of points[e]	1061 of 1331
Number of components[f]	8
PRESS[g]	0.60
Cross-validated[h] q^2	0.74

[a] The *x*, *y* and *z* dimensions of the grid sampled by the probe atom, centred about the centre of mass of the template compound.
[b] The separation between lattice points of the grid.
[c] The maximum energy allowed for any particular probe–substrate interaction.
[d] If the variance of the probe–substrate interactions at a particular lattice point for all of the members of the training set was below this value, that point was excluded from the data set.
[e] The number of lattice points remaining after filtering out points with low variance – the second value is the total number of points sampled prior to filtering.
[f] The optimal number of latent variables (PLS components) included.
[g] Predicted residual sum of squares [25].
[h] Eq. (1) [25].

is illustrated in Fig. 3a, which shows all of the compounds aligned in the extended conformation.

The values of the CoMFA parameters used in the best model for predicting enantiomeric excess are shown in Table 2. The optimal parameters for our model were determined by choosing those that maximized the ability to predict e.e. as measured by the cross-validated q^2. The final model created to predict the e.e. for sulfide oxidation by *R. erythropolis* had a q^2 of 0.74, which suggests that it is of high quality. The most influential of the various parameters listed in Table 2 on the value of q^2 were the substrate conformation and biological activity that was modelled. The models were not very sensitive to moderate changes of the values of the remaining parameters (see discussion in Section 4). Fig. 3b depicts a head-on view of a portion of the selected CoMFA grid, in order to depict the positioning of the probe atom about the aligned substrates of the training set.

This best CoMFA model for predicting e.e. was tested with an external test set of substrates (Table 3). This test set consisted of five of the compounds from Holland et al. [9] that had not been included in the training set (compounds 8, 18, 23, 24, and 34), but are members of the same classes of compounds. In addition, in order to examine the flexibility of the model, two amino acid sulfides were tested (compounds 60 and 61, Table 3). The high predictive ability of the model is illustrated both by the small residuals listed in Table 3 and by Fig. 4, which plots the predicted versus experimentally determined e.e. values for both the training set (Table 1) and the external test set (Table 3) compounds. The ability of the model to make relatively accurate predictions of the e.e. for a class of compounds not considered in the training set (the amino acid sulfides) demonstrates its potential as a predictive tool.

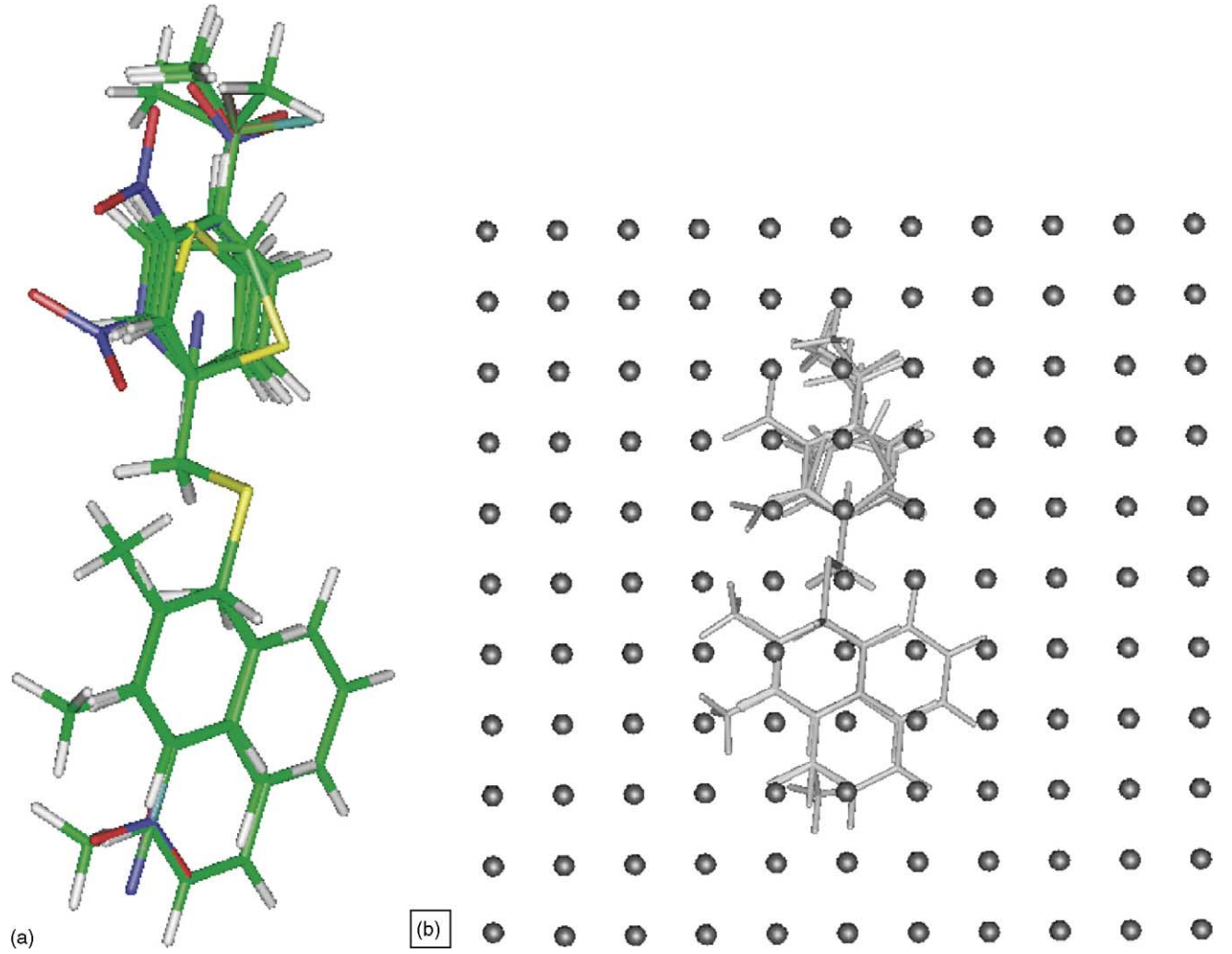

Fig. 3. (a) Representation of the 26 members of the training set aligned to the template compound in the extended conformation. S: yellow; C: green; O: red; N: blue; F: cyan; Br; brown; Cl: light green; I: purple. (b) Aligned training set with portion of surrounded lattice grid used for optimized CoMFA model; lattice spacing of 2 Å. Figures prepared using InsightII [30].

Also plotted in Fig. 4 are the predicted versus experimentally determined e.e. values for the four compounds identified as outliers (Table 1: compounds 38, 51, 41 and 45). The large residuals for these four compounds underline the inadequacy of the model for predicting e.e. for these four compounds, and presumably others of their classes.

The q^2 values for optimized CoMFA models for the prediction of enantiomeric excess given the two different substrate conformations examined are compared in Table 4. The CoMFA model for predicting e.e., obtained where all substrates were aligned in the extended conformation, is significantly more predictive than that obtained when the substrates

Table 3
External test set

Compound number[a]	Formula	Experimental e.e.[a]	Predicted e.e.	Residual[b]
8	*p*-$BrPhSCH_3$	0.76	0.72	−0.04
18	*p*-$C_2H_5PhCH_2SCH_3$	0.53	0.52	−0.01
23	*o*-$CH_3OPhCH_2SCH_3$	0.44	0.42	−0.02
24	*p*-$CH_3OCH_2PhCH_2SCH_3$	0.61	0.48	−0.13
34	*p*-$CNPhCH_2SCH_3$	0.72	0.84	0.12
60	*N*-MOC-L-methionine methyl ester	0.83	0.80	−0.03
61	*N*-MOC-D-methionine methyl ester	0.96	0.99	0.03

[a] Compound numbers and e.e. values correspond to those reported in Tables 1, 2 and 4 in Holland et al. [9].
[b] Residual = predicted e.e. − experimental e.e.

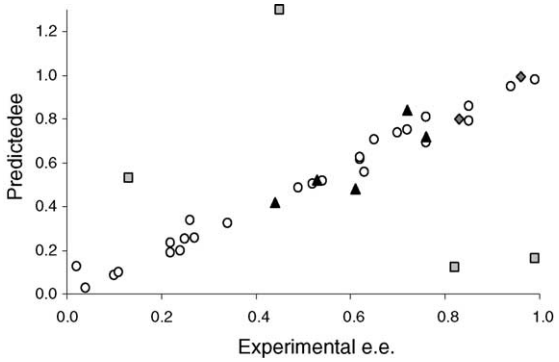

Fig. 4. Predicted vs. experimentally determined e.e. values for the sulfur oxidation reaction of the enzyme DszC of *R. erythropolis*. The open circles are the values for the training set compounds (Table 1), while the grey squares represent the values of the four compounds identified as outliers (Table 1). The solid triangles are the values for the external test set (Table 3), while the grey diamonds represent the two amino acid sulfides tested (Table 3).

were aligned in the folded conformation (a cross-validated q^2 of 0.74 versus 0.28). The high q^2 value obtained for the extended conformation model (Table 4, row 1) corresponds to a high level of agreement between the actual and predicted values of the enantiomeric excess. The apparent success in predicting the e.e. for the extended model suggests that the substrate, when bound in the DszC active site, is more likely to assume an extended conformation than a folded conformation.

We also attempted to create CoMFA models using the product yield as the measure of biological activity. The data for yield were taken from Holland et al. [9], where yield measures the mass fraction of sulfoxide product to sulfide starting material. None of these models were at all successful at predicting product yield (the best values of q^2 for models developed using the extended and folded conformations of the substrate were −0.07 and 0.19, respectively). In retrospect, this is not a surprising since the data were obtained for a whole cell biocatalyst; both inefficient substrate transport through cell membranes and possible interactions with other cellular components would decrease the experimental yield. The present CoMFA methodology assumes uninhibited transport to the active site and cannot hope to capture the intricacies of substrate transport, although one crude approximation would be to include an additional descriptor such as log *P* to attempt to account for this phenomenon.

Figs. 5 and 6 are contour diagrams that illustrate regions of high positive and negative correlation to the predicted

Table 4
Ability of CoMFA model to predict enatiomeric excess for different substrate conformations[a]

Conformation	Activity modelled	Cross-validated q^2
Extended	e.e. only	0.74
Folded	e.e. only	0.28

[a] These models were created using all 26 members of the training set and employed the optimized parameters shown in Table 2 unless indicated otherwise.

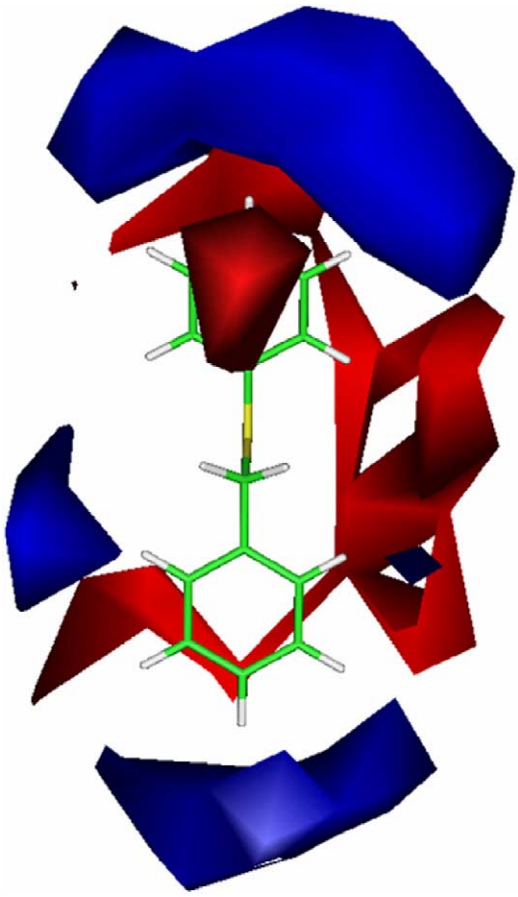

Fig. 5. Top-view contour diagram of the correlation between predictive ability for enantiomeric excess and regions in space in the putative binding pocket of the enzyme DszC using both steric and electrostatic interactions to model this relationship. The blue represents regions of positive correlation (enhanced e.e.), while the red represents regions of negative correlation (diminished e.e.). The template compound, $PhCH_2SPh$, is shown for perspective. Figure prepared using InsightII [30].

enantiomeric excess as predicted by the CoMFA model with the parameters described in Table 2. The contribution of each three-dimensional lattice point to the predicted e.e. is computed from the corresponding CoMFA coefficient. The contour diagrams are a visual representation of important interactions that may exist within the active site of the enzyme. The blue regions are positively correlated, while the red regions are negatively correlated to e.e. of the sulfoxide product. That is to say, the presence of substituents on the substrate that project into the regions coloured blue are associated with enhanced e.e., whereas substituents in the regions coloured red are associated with diminished e.e. of the sulfoxide product. These diagrams concur with the proposal by Holland et al. [9] that *para*-substituents on both aromatics, particularly the α-phenyl, lead to products with higher enantiomeric excesses that those with *ortho*- or *meta*-substitutions. The positive correlations in the *para*-position is obvious, while the regions of negative correlation illustrate the sensitivity of the enzyme to *ortho*- and *meta*-substitutions on both the α and β rings.

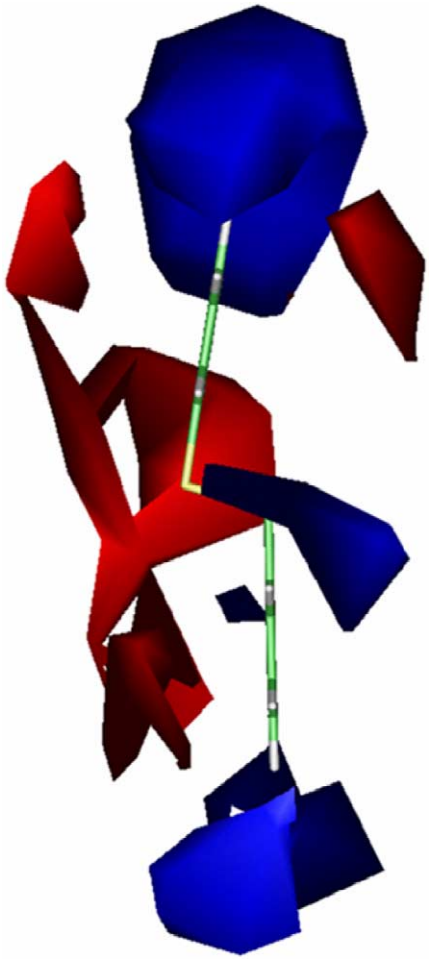

Fig. 6. Side-view contour diagram of the correlation between predictive ability for enantiomeric excess and regions in space in the putative binding pocket of the enzyme DszC using both steric and electrostatic interactions to model this relationship. The blue represents regions of positive correlation (enhanced e.e.), while the red represents regions of negative correlation (diminished e.e.). The template compound, $PhCH_2SPh$, is shown for perspective. Figure prepared using InsightII [30].

The asymmetry of the modelled binding pocket is quite obvious in Figs. 5 and 6. Interactions that are significant for producing the R enantiomer in excess occur predominantly on three sides when visualized in a 'top-down' fashion (Fig. 5). One possible explanation for the 'open' region seen in the lower left side of Fig. 5 is the presence of an additional pocket within the binding site. Such an explanation could account for the poor degree to which the biphenyl compounds (41 and 45 in Table 1) were predicted by this model, if this pocket provided an alternative binding mode for these types of compounds. Instances of other sulfur oxygenases that bind substrates in one of two different modes have been reported in the literature [21].

Figs. 5 and 6 represent the final CoMFA model that incorporates both steric and electrostatic components of the interaction energy of the probe with the training set of substrates. Table 5 shows the effect upon predictive ability (q^2) of the model when considering only electrostatic or van der Waals potential energies, as compared to the total energy. The electrostatic interactions clearly have more of an impact on the outcome of the prediction (the q^2 using electrostatic energy only is 0.54 versus 0.16 for van der Waals only). A contour diagram of the model created that employs only the electrostatic interactions is shown in Fig. 7; the similar-

Table 5
Comparison of energy terms employed[a]

Energy term(s)	Number of components	PRESS	Cross-validated q^2
Electrostatic only	2	0.98	0.54
van der Waals only	3	1.80	0.16
Total (elect. + vdW)	8	0.60	0.74

[a] These models were created using all 26 members of the training set and employed the optimized parameters shown in Table 2 unless indicated otherwise.

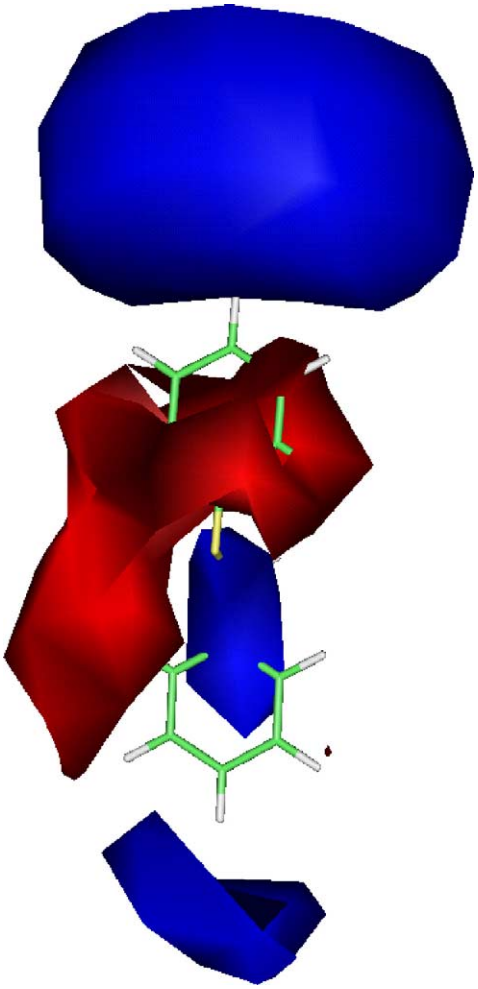

Fig. 7. Top-view contour diagram of the correlation between predictive ability for enantiomeric excess and regions in space in the putative binding pocket of the enzyme DszC using only the electrostatic interactions to model this relationship. The blue represents regions of positive correlation (enhanced e.e.), while the red represents regions of negative correlation (diminished e.e.). The template compound, $PhCH_2SPh$, is shown for perspective. Figure prepared using InsightII [30].

ity to Figs. 5 and 6 is obvious. The regions of high positive correlation in this diagram correspond to regions where increased electron density is favourable to large e.e. values for sulfoxide product (the probe used to model the interactions had a positive charge). This data corroborates the observation in [9] that substrates with higher electron-withdrawing groups (e.g., *para*-NO_2, and -CN) provided higher e.e. values.

3. Conclusion

In this paper we have outlined the construction of a highly predictive CoMFA model for the enantiomeric excess of the sulfur oxidation for a broad range of products catalyzed by the mono-oxygenase, DszC, of the bacterium *R. erythropolis* IGTS8. Although we found that the yield was not well modelled, a CoMFA model was able to predict, with a q^2 of 0.74, the enantiomeric excess for a wide range of substrates of this reaction. Comparison of two low energy conformations suggests that the most likely bioactive form for the classes of substrates for which the model is predictive has an extended conformation. Given the high predictive ability of the best model, it was surprising to find that the dibenzyl compounds, which included the template molecule, were not well predicted in concert with the other classes of substrates. This suggests the possibility of a different binding mode for this type of compound. Contour diagrams constructed from the CoMFA model illustrate regions of important steric and electrostatic interactions within the binding pocket of the enzyme and are consistent with previous conclusions made by Holland et al. [9] about the effect on enantioselectivity of various substitutions on the sulfide substrate.

4. Experimental

Biological activity data was obtained from Holland et al. [9] for the biotransformation of sulfides by *R. erythropolis* IGTS8 BKO-53. Yields were reported as a w/w fraction of sulfoxide product to sulfide starting material while enantiomeric excess was reported as a fraction of R enantiomer of the total weight of product.

The training set was chosen from among the 68 compounds for which data was available (Tables 1–5 in [9]). Compounds were selected in order to maximize the structural variation of the members and to provide a broad range of activity values. Only those substrates that had fewer than 10 energetically non-degenerate conformational minima were considered for the training set. A subset of 30 substrates representing four classes of compounds (substituted methyl phenyl sulfides, substituted benzyl sulfides, dibenzyl sulfides, and heterocyclic sulfides) was initially chosen for the training set. Examination of preliminary models showed that four of these compounds were much more poorly predicted than any of the other compounds. Further investigation found that these compounds belonged to two groups that were unique to the training set. Two of these compounds (compounds 38 and 51 from Tables 2 and 4 in [9]) were the only members of the training set that had been produced predominantly as the S enantiomer. The other two compounds (compounds 41 and 45 from Table 3 in [9]) were the only two members of the training set that were dibenzyl sulfides. These four compounds were omitted and the remaining 26 compounds (Table 1) were used to construct the final model. These compounds had yields ranging from 4 to 95% and enantiomeric excess values ranging from 2 to 99%.

The low energy conformations of each of the members of the training set were determined using the SPARTAN molecular modelling package [29]. The compound identified as the 'ideal substrate' in Ref. [9], benzyl mercapto-benzene ($PhCH_2SPh$), was used as the template molecule. This compound was chosen because of its high activity (e.e. of 0.99) and its relative lack of conformational flexibility ($PhCH_2SPh$ assumes only two low-energy conformations – see Fig. 2). The molecular modelling package InsightII [30] was used to align individually the members of the training set to the template. The alignment was carried out by rigid superposition of the substrate and template, where the root-mean-square distance between all pairs of corresponding heavy atoms (C, N, O, and S) was minimized. The closest matching conformations of each member of the training set were aligned individually to both the extended and folded conformations of the template molecule shown in Fig. 2.

Using the CHARMm forcefield [31,32], the interaction energies of a probe atom with each of the members of the training set were calculated within the confines of a three-dimensional lattice. The selected probe, an sp^3 carbon atom with a +1 charge, was positioned at each point in the lattice and the electrostatic, van der Waals and total (electrostatic + van der Waals) interaction energies were calculated. A distance-dependent dielectric constant was employed for the electrostatic interaction calculations.

An in-house partial least-squares programme, based upon the NIPALS algorithm [33], was used to develop all CoMFA models. The 'leave-one-out' cross-validation method [26,34] was used to determine the optimal number of latent variables. The optimal number of latent variables was that which provided the lowest cross-validated predicted residual sum of squares (PRESS). The resultant CoMFA model was then created using this optimal set of latent variables with the data for the entire training set. The cross-validated q^2, used as a measure of predictive ability, was calculated as shown below [25]:

$$q^2 = \frac{\mathrm{SD} - \mathrm{PRESS}}{\mathrm{SD}} \quad (1)$$

The PRESS is the predicted residual sum of squares and is the sum, over all substrates, of the squared differences between

the actual and predicted biological activity values, whereas SD is the sum of the squared deviations of the actual activity values from their mean. Values of q^2 close to one are highly predictive, while values of zero or lower are no better than random predictions [25].

Several CoMFA models were constructed to examine the effects of varying certain standard parameters. The effect of changing the grid dimensions, step size, energy cutoff, and variance cutoff were all probed by iteratively creating models that varied in only one of these parameters (see footnotes to Table 2 for explication of these parameters). Varying the energy cutoff from 10 to 30 kcal mol^{-1} and the variance cutoff from 1 to 5 (kcal mol^{-1})^2 did not have a significant effect on the predictive ability of the model, as the value of q^2 changed by at most 10%. Increasing the overall dimensions of the cubic grid beyond ±10 Å about the aligned compounds did not improve q^2, as the grid points at the extremities of the box were eliminated due to the lack of variation in the interaction energies at these locations. A smaller cubic grid of only ±5 Å about the aligned compounds was not sufficiently large to incorporate all important regions of interaction space, as the value of q^2 dropped by up to 25%. Increasing the separation between lattice points on the grid from 2 to 3 Å decreased q^2 only by about 10%. Due to computer memory limitations, we were unable to work with grids of sufficiently large dimensions to enclose the aligned compounds and having a lattice spacing of much less than 2 Å. In addition, CoMFA models were created for both the extended and folded conformations to determine which provided the best predictive ability. The suitability of both the yield and enantiomeric excess as measures of biological activity were also examined.

A test set of compounds was chosen from among the remaining 42 compounds in Ref. [9] that were not employed in the training set. The e.e. values of this test set (Table 3 and Fig. 4) were predicted using the best CoMFA model and employed the optimal parameter set (Table 2).

The coefficients of the best CoMFA models were used to create contour plots of the correlation between the energy fields (either steric plus electrostatic interactions or electrostatic interactions only) and the enantiomeric excess (Figs. 5–7). InsightII [30] was used to create the contours by joining points within the lattice that had CoMFA coefficients that were similar in magnitude. Colour was used to distinguish between positive (blue) and negative (red) correlations of the lattice point to the biological activity. If a lattice point is positively correlated with e.e., the presence of a substituent on the substrate molecule in this location enhances the e.e. of the sulfoxide product.

Acknowledgements

This paper is dedicated to the memory of Herbert L. Holland. The corresponding authors would like to acknowledge the assistance of F. Brown and helpful suggestions of S.M. Rothstein. This work was supported by grants from the Natural Sciences and Engineering Research Council of Canada (NSERC).

Appendix A. Availability of supporting information

Protein databank files of the aligned compounds listed in Tables 1 and 3, used to develop the CoMFA model presented here, are available by contacting H.L. Gordon.

References

[1] W.R. Finnerty, Ann. Rev. Microbiol. (1992) 46.
[2] K.A. Gray, O.S. Pogrebinsky, G.T. Mrachko, L. Xi, D.J. Monticello, C.H. Squires, Nat. Biotechnol. 14 (1996) 1705.
[3] C. Oldfield, N.T. Wood, S.C. Gilbert, F.D. Murray, F.R. Faure, Ant. van Leeuwen 74 (1998) 119.
[4] B. Mcfarland, Curr. Opin. Microbiol. 2 (1999) 257.
[5] B. Lei, J. Bacteriol. 178 (1996) 5699.
[6] M.Z. Li, C.H. Squires, D.J. Monticello, J.D. Childs, J. Bacteriol. 178 (1996) 6409.
[7] S.A. Denome, C. Oldfield, L.J. Nash, K.D. Young, J. Bacteriol. 176 (1994) 6707.
[8] C. Oldfield, O. Pogrebinsky, J. Simmonds, E.S. Olson, C.F. Kulpa, Microbiology 143 (1997) 2961.
[9] H.L. Holland, F.M. Brown, A. Kerridge, P. Pienkos, J. Arensdor, J. Mol. Catal. B: Enzym. 22 (2003) 219.
[10] H.L. Holland, F.M. Brown, F. Barrett, J. French, D.V. Johnson, J. Ind. Microbiol. Biotechnol. 30 (2003) 292.
[11] W. Komatsu, Y. Miura, K. Yagasaki, Lipids 33 (1998) 499.
[12] M.Y. Brusniak, R.S. Pearlman, K.A. Neve, R.E. Wilcox, J. Med. Chem. 39 (1996) 850.
[13] G. Solladie, Synthese 3 (1981) 185.
[14] S. Nakamura, Y. Watanabe, T. Toru, J. Org. Chem. 65 (2000) 1758.
[15] M.C. Carreno, Chem. Rev. 95 (1995) 1717.
[16] H.L. Holland, Organic Synthesis with Oxidative Enzymes, VCH Publishers, New York, 1992.
[17] S. Patai, Z. Rappaport, C. Sterling, The Chemistry of Sulphones and Sulphoxides, Wiley, New York, 1988.
[18] M.A.M. Capozzi, C. Cardellicchio, G. Fracchiolla, F. Naso, P. Tortorella, J. Am. Chem. Soc. 121 (1999) 4708.
[19] M.I. Donnoli, S. Superchi, C. Rosini, J. Org. Chem. 63 (1998) 9392.
[20] H.L. Holland, F.M. Brown, D. Lozada, B. Mayne, W.R. Szerminski, A.J. van Viet, Can. J. Chem. 80 (2002) 633.
[21] H.L. Holland, Nat. Prod. Rep. 18 (2001) 171.
[22] A. Kerridge, A. Willetts, H.L. Holland, J. Mol. Catal. B: Enzym. 6 (1999) 59.
[23] H.L. Holland, C.D. Turner, P.R. Andreana, D. Nguyen, Can. J. Chem. 77 (1999) 463.
[24] H.L. Holland, F.M. Brown, G. Lakshmaiah, B.G. Larsen, M. Patel, Tetrahedron: Asymmetr. 8 (1997) 683.
[25] R.D. Cramer, D.E. Patterson, J.D. Bunce, J. Am. Chem. Soc. 110 (1988) 5959.
[26] R.D. Cramer, S.A. DePriest, D.E. Patterson, P. Hecht, in: H. Kubinyi (Ed.), 3D QSAR in Drug Design: Theory, Methods and Applications, ESCOM, Leiden, 1993, p. 443.
[27] U. Norinder, Perspect. Drug Design 12 (1988) 25.

[28] S.A. DePriest, D. Mayer, C.B. Naylor, G.R. Marshall, J. Am. Chem. Soc. 115 (1993) 5372.
[29] Wavefunction Incorporated, Spartan Open GL version 5.1.3, Irvine, California, 1998.
[30] Accelrys Inc., InsightII, Release I2000, Accelrys Inc., San Diego, 2000.
[31] B.R. Brooks, R.E. Bruccoleri, B.D. Olafson, D.J. States, S. Swaminathan, M. Karplus, J. Comput. Chem. 4 (1983) 187.
[32] Accelrys Inc., CHARMm®, Accelrys Inc., San Diego, 2001.
[33] F. Lindgren, P. Geladi, S. Wold, J. Chemomet. 7 (1993) 45.
[34] S. Wold, Technometrics 20 (1978) 397.