9330 *Biochemistry* **2009,** *48,* 9330–9339
DOI: 10.1021/bi901168q

BIOCHEMISTRY
including biophysical chemistry & molecular biology
Current Topic

Identification of Emerging Quasi-Species in Directed Enzyme Evolution[†]

Sanela Kurtovic[‡] and Bengt Mannervik*

Department of Biochemistry and Organic Chemistry, Uppsala University, BMC, Box 576, SE-75123 Uppsala, Sweden [‡]*Current address: Atherosclerosis Research Unit, Department of Medicine, Karolinska Institutet, Center for Molecular Medicine L8:03, Karolinska University Hospital Solna, SE-17176 Stockholm, Sweden.*

Received July 9, 2009; Revised Manuscript Received September 10, 2009

ABSTRACT: The bases of enzyme evolution are structural changes in protein scaffolds combined with recognition and propagation of novel variants with valuable functional properties. Structural diversification may be accomplished by a variety of methods, including random mutations, homologous recombinations, and insertions and deletions of coding DNA sequences. The functional consequences of mutations are manifested at the protein level and are dependent on a substrate matrix, when catalytic properties are requested. Libraries of variant enzymes showing promiscuous activities can be interrogated with a set of alternative substrates. We demonstrate using a library of glutathione transferases (GSTs) that the functional properties are not uniformly distributed in substrate−activity space but form clusters, or quasi-species. Multivariate analysis facilitates the identification of such quasi-species, which can be regarded as the proper developing units in molecular evolution.

The natural plasticity of the polypeptide structure gives rise to proteins that serve an essentially unlimited number of functions in living cells. Directed evolution of proteins is a well-established approach to exploiting this inherent potential and thereby creating novel devices for a variety of applications (*1*). Particular attention has been directed to the engineering of enzymes with catalytic properties tailored for requirements in biotechnology. In this review, we will focus on our own approach to multivariate catalytic activities, but the analytical methods described are generally applicable to evolving entities, other than enzymes, for which quantitative functional data can be obtained.

In general, directed enzyme evolution can be divided into three aspects: (i) generation of structural diversity in a protein scaffold, (ii) acquisition of functional information, and (iii) analysis of the collected data. Diversity is usually accomplished iteratively by mutagenesis applied to an existing molecular structure. This redesign requires optimization to yield maximal variability without compromising the structure and stability of the protein scaffold. Numerous methods for producing mutant libraries by stochastic mutagenesis have been described (*2*−*4*). Alternatively, structure-based targeting of defined positions in the polypeptide chain can give structurally focused randomization of key residues by site-directed mutagenesis (*5*) or whole-gene synthesis. To find enzyme mutants with desired properties, one must query the mutant library through selection or screening (*6*−*8*). In other words, the enzyme activity should be tested against relevant substrates. Selection can be very powerful if the targeted activity can offer a proliferative advantage to cells expressing the enzyme. However, most enzyme reactions cannot function as selectable traits in a host cell. The display of mutants on phage, ribosomes, or mRNA is an efficient selection tool for binding proteins (*9*) but is generally less useful for selection of enzyme activities. In most cases, the analysis of a mutant library will therefore be based on screening, in which mutants from the library are individually characterized. In the screening, a variety of functions can be examined in parallel, to provide a valuable information matrix. The use of small libraries allows analysis of all mutants present, but with large mutant libraries, only a limited sampling of the population is feasible.

The evolution of enzyme function is a multidimensional issue. The selection of a favorable activity has to be balanced against disadvantageous functions and to secure suitable physical properties such as foldability and stability, as well as compatibility with other components in the milieu in which the enzyme should be active. In general, enzyme evolution is consequently an optimization problem. In this review, we illustrate how multivariate data analysis can be used as a tool to monitor evolutionary trajectories in functional space and facilitate directed evolution of enzyme functions. Furthermore, novel concepts regarding enzyme functionalities in directed evolution will be discussed. The examples given will be restricted to the multifunctional glutathione transferases (GSTs), which display activities with alternative substrates. However, the approach can include any property relevant to enzyme evolution that can be quantified in a comparable manner.

OUTLINE OF OUR EXAMPLE

Figure 1 depicts the design of an investigation that we will use as a paradigm for our review. The top part of the figure describes the generation of a library of variant enzymes by a stochastic approach. The key point is the creation of structural variability, which is essential to the emergence of novel functionalities. In this case, we will explore the functional consequences of family shuffling (*10*) of DNA encoding six mammalian GSTs. However, other methods for introducing stochastic mutations into one or several parental sequences can also be used. The next step is to express the variant proteins for functional characterization.

[†]The authors' research has been supported by the Swedish Research Council and the Swedish Cancer Society.

*To whom correspondence should be addressed. Phone: +46-18 471 4539. Fax: +46-18 558431. E-mail: Bengt.Mannervik@biorg.uu.se.

 Published on Web 09/11/2009

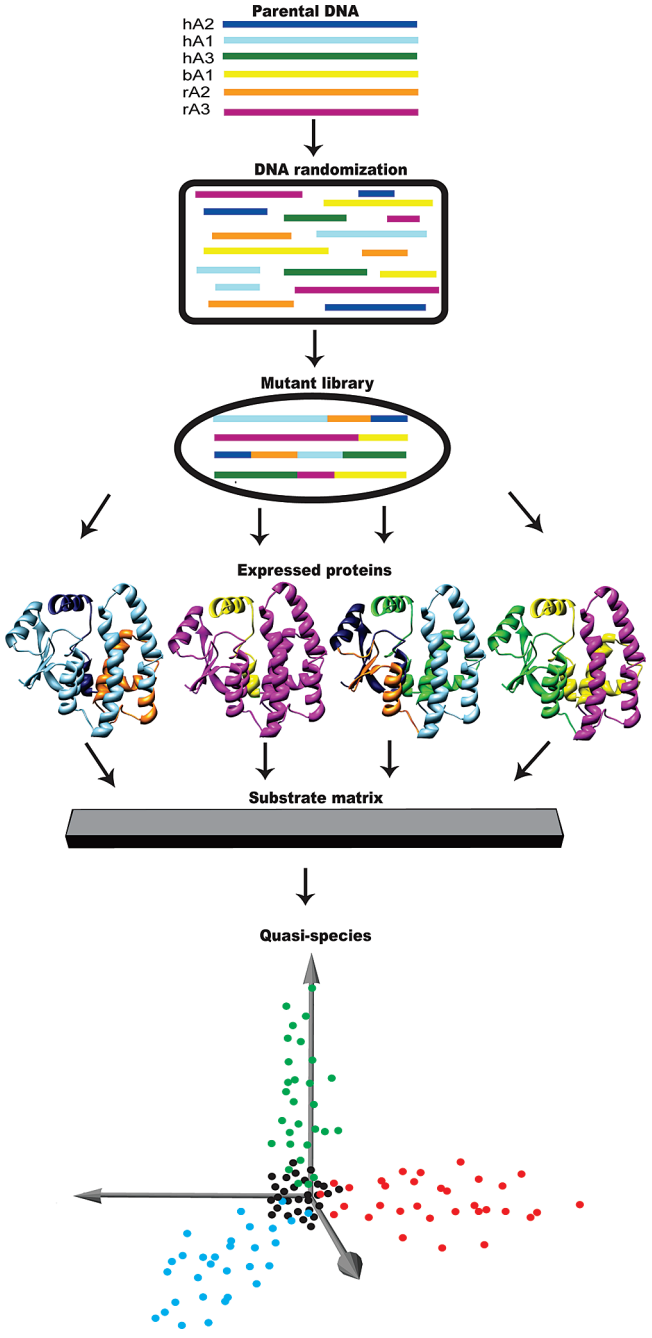

FIGURE 1: Outline of an experimental system involving a mutant library of recombinant enzymes and screening of enzyme activities with alternative substrates, followed by multivariate analysis and identification of functional quasi-species. In our example, DNA sequences encoding six parental GSTs (hA1, hA2, hA3, bA1, rA2, and rA3) of the alpha class were randomly cleaved with DNase I into DNA fragments of different sizes. These DNA fragments were subsequently recombined and heterologously expressed to generate full-length variant enzymes (*11*). Chimeric proteins generated in this manner fold into three-dimensional structures in which different parental segments of diverse lengths are the building blocks. The catalytic functions of randomly sampled enzyme variants are cross-examined, via a "substrate matrix" composed of alternative substrates. Multivariate analysis of the activity data allows the GST variants to be classified into quasi-species. The GSTs are abbreviated as follows: hA1, human GST A1-1; hA2, human GST A2-2; hA3, human GST A3-3; bA1, bovine GST A1-1; rA2, rat GST A2-2; rA3, rat GST A3-3 (modified from ref *29*).

For our analysis, the variant GSTs were expressed as recombinant proteins via heterologous expression in *Escherichia coli* (*11*). The following step involves determination of catalytic activities of the different enzyme variants. The library was estimated to contain 10^6 mutants, and the screening was restricted to sparse sampling of individual clones of bacteria for analysis. In our example, enzyme variants were tested at random, but selection involving phage-displayed GSTs has also been used (*12*, *13*). When applicable, phenotypic selection can be very powerful and allow the mining of very large mutant libraries. On the other hand, screening, which is generally limited to smaller libraries, offers the advantage of generality and, most importantly, the parallel examination of multiple functional properties.

SUBSTRATE MATRIX

We define the substrate matrix as the array of substrates with which enzyme activity is tested. The substrate matrix forms the basis for the functional distinction of a given enzyme from other enzymes. Traditionally, enzymes have been considered to catalyze one particular chemical reaction, such as the hydrolysis of urea by the enzyme urease. If no compound in the substrate matrix other than urea can serve as a substrate for urease, the activity with urea is clearly orthogonal to the activities of all substrates of the matrix that may be catalyzed by other enzymes, provided that these enzymes are not active with urea. However, the concept of orthogonality has less value for distinction among enzymes that catalyze reactions with alternative substrates of which at least one substrate is shared among the enzymes analyzed. As we will demonstrate below, the substrate matrix is instrumental in the more refined distinction and characterization of enzymes.

Promiscuity in the sense of the ability to catalyze reactions with several alternative substrates appears to be relatively common when the substrate matrix is expanded to include more than the conventional natural substrates (*14*). However, unnatural alternative substrates may have limited interest in a biological setting, except from a toxicological point of view. On the other hand, in the protein engineering of enzymes for novel biotechnical applications, unnatural substrates can be both relevant and useful.

ALTERNATIVE SUBSTRATES

Figure 2 shows the array of substrates used in our example of analysis. In this case, the reactive sulfur of the second substrate glutathione will undergo a variety of alternative chemical reactions such as alkylation, arylation, nitrosylation, acylation, and thiocarbamoylation (*15*). In addition, glutathione serves as a base in a steroid double-bond isomerization. We will consider enzyme activities determined under standard conditions for each substrate, but in other cases, the substrate matrix can be expanded to include ranges of substrate concentrations that are relevant to the milieu in which the enzyme should be active. At substrate concentrations sufficiently below the K_m value for the tested substrates, the catalytic activities are proportional to k_{cat}/K_m, which can be used as a quantitative measure of discrimination between alternative substrates (*16*). The use of several substrates is obviously a prerequisite for determining functional versatility and substrate selectivities of the enzymes under investigation.

ENZYMES AS VECTORS IN SUBSTRATE−ACTIVITY SPACE

By definition, enzymes have catalytic activity. This functional property can be represented by a vector, and the length of the vector is a measure of the magnitude of the catalytic function. Enzymes that catalyze reactions with alternative substrates form

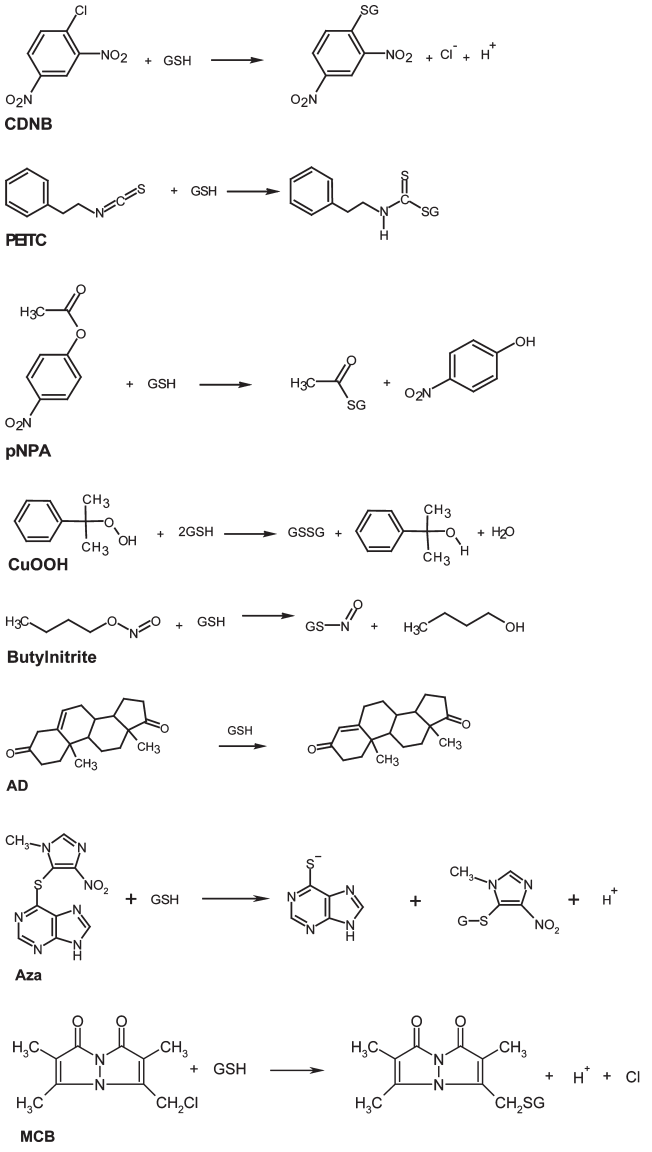

FIGURE 2: Substrate matrix encompassing GST-catalyzed reactions with alternative substrates representing different types of chemical transformations: arylation [1-chloro-2,4-dinitrobenzene (CDNB) and azathioprine (Aza)], alkylation [monochlorobimane (MCB)], addition to isothiocyanate [phenethyl isothiocyanate (PEITC)], transacylation [*p*-nitrophenyl acetate (pNPA)], hydroperoxide reduction [cumene hydroperoxide (CuOOH)], transnitrosylation (butylnitrite), and steroid isomerization [Δ^5-androstene-3,17-dione (AD)].

vectors in multidimensional substrate−activity space. A vector that falls on one of the substrate−activity axes with no contribution of any alternative substrate has the highest substrate selectivity possible and is traditionally considered to demonstrate "absolute substrate specificity". A vector with contributions of activities with several substrates displays broad substrate acceptance, also referred to as substrate "ambiguity" (*17*) or "promiscuity"(*18*). Comparison of the catalytic properties of two enzymes can be made by comparison of their respective vectors in substrate−activity space. Vectors with the same direction are obviously functionally the same in terms of substrate acceptance and are therefore considered as the same enzyme, even if the lengths of the vectors differ. The length of a vector measured per volume, or mass, of a biological sample will depend on the amount of catalytically active enzyme. When assayed under the same conditions, different purified samples of an enzyme are expected to have the same activity, if they have the same primary structure. However, differences may appear among allelic variants or isoenzymes in the same organisms and can also be caused by post-translational modifications and other processes. In contrast, enzymes represented by activity vectors with distinctly different directions can be regarded as different enzymes.

COMPARISON OF ENZYME VECTORS AND THE EMERGENCE OF NEW ENZYMES

In the course of enzyme evolution, in particular when enzyme variants are active with several alternative substrates, the question of whether a new enzyme has emerged from the progenitor may arise. For this question to be answered, the emerging variant has to be compared with at least one parental enzyme. We are focusing our discussion on functional properties, and if the directions of the vectors differ significantly, the progenitor and the variant offspring will be considered as separate enzymes. The direction rather than the length of a vector is the salient feature. Experimental variance will probably cause some divergence of two enzyme vectors, even if they represent the same enzyme. Multivariate statistical tests based on estimates of the experimental variance can be used to test the likelihood that the vectors have different directions in substrate−activity space (*19*). However, a more important query is how much intrinsic functional divergence caused by structure-based differences can be accepted before a new enzyme has to be invoked. This issue becomes prominent in the comparison of variant enzymes with broad substrate acceptance. To address this question, it may be necessary to examine the catalytic properties of a set of structural variants of one and the same parental enzyme and compare the typical properties of this group with the properties of an emerging novel enzyme. The direction of the novel enzyme vector is expected to deviate significantly from that of the vectors of the parental variants.

FUNCTIONAL QUASI-SPECIES

In a biological population, a given enzyme may naturally occur in variant forms with minor differences in their catalytic properties. The variants would still be regarded as the same enzyme. In the directed evolution of enzyme functions, comparable ensembles of variants with similar properties will appear. Represented in substrate−activity space, the variants form a multidimensional swarm of data points, which can also be visualized as a bundle of vectors. Such ensembles of enzyme variants may be considered to form molecular quasi-species defined by common functions. The quasi-species is a virtual entity representing a group of variants (*20*).

EMERGENCE OF CATALYTIC ACTIVITY IN A PROTEIN SCAFFOLD UNDERGOING STOCHASTIC MUTATIONS

Figure 3 exemplifies scatter plots of catalytic activities measured separately with two alternative substrates, 1-chloro-2,4-dinitrobenzene (CDNB) and Δ^5-androsten-3,17-dione (AD). The *x*-axis shows the serial numbers of the mutants analyzed, but the numbers can also be considered a time series indicating the emergence of different mutants over the course of an evolutionary period. The important aspect is the unsystematic expression of different levels of catalytic activity with a given substrate. The activities determined with CDNB appear to have a random dispersion of amplitude along the *x*-axis. By contrast, the

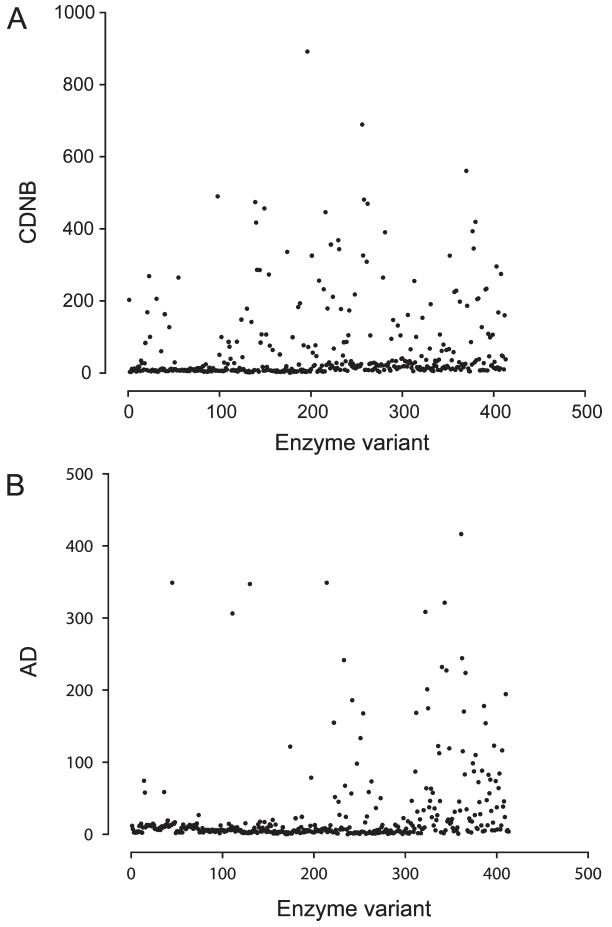

FIGURE 3: Scatter plots of catalytic activities determined in lysates of bacteria expressing GST variants randomly picked from clones in the recombinant mutant library. Mean values of duplicate measurements in lysates are used throughout this work for all analyses in this paper: (A) activities with CDNB as a substrate plotted vs clone number, showing random scattering of high- and low-activity mutants, and (B) activities with AD as a substrate. Activities are given in arbitrary units. In this series of data, the second half, which had been assayed with AD on a later occasion, demonstrated excessive scattering in comparison with the first half. To avoid differences in the experimental variance as a possible confounding factor, the second half of the data set (numbers 265−456) was excluded in the subsequent multivariate data analysis presented in this review (modified from ref *29*).

activities with AD appear to be divided into two halves along the same axis and scatter in two apparently different manners. In a naturally evolving system, such a change in the distribution with time may signify the introduction of a new level of structural variation, or changes in the milieu in which the enzyme variants express their activity. However, in this case, the first half of the data set was assayed first and the second half on a later occasion. For multivariate analysis, it is important to examine the data set for such deviations from randomness. Otherwise, bias introduced in the experimental measurements can improperly be interpreted as clustering of data due to the intrinsic functional variability in the mutant library. In the analysis presented in this review, we have therefore restricted the data set to measurements conducted in the first half of the investigation, to avoid confounding factors. An alternative would have been to label the data points such that they can be traced back to the two halves of the data set assembled on different occasions.

It can be noted that, in this case, analysis of the complete data set has been performed and given basically the same clustering as described below. However, the formation of the clusters was

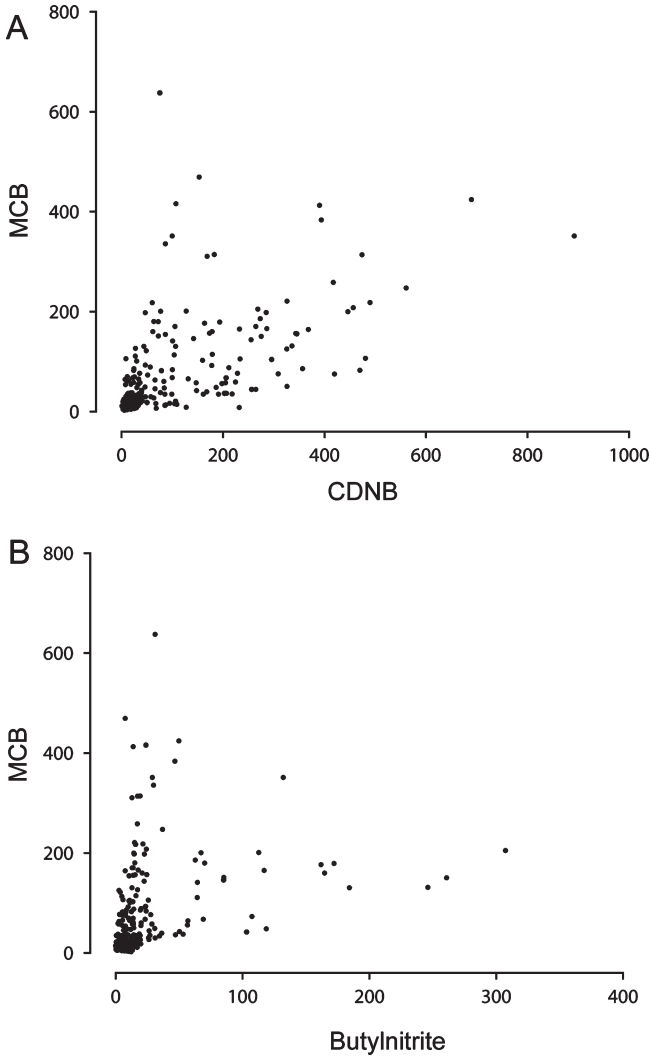

FIGURE 4: Two-dimensional scatter plot demonstrating different associations of GST activities with alternative substrates. Each GST variant was assayed with all eight of the alternative substrates in Figure 2, and activities with alternative substrates were plotted pairwise. (A) Correlation between MCB and CDNB activities without marked divergence into more than one cluster. (B) Correlation between MCB and butylnitrite activities indicating two diverging distributions among the GST variants. Activities are given in arbitrary units.

strongly influenced by the activity measured with AD, and the apparent bimodal distribution of the AD activities (Figure 3B) will influence the location of individual points in the various plots.

Figure 4 shows alternative two-dimensional representations of the data set, in which the activities with two alternative substrates are plotted against one another. In panel A, the alternative activities of the GST variants sampled are spread in the plain of CDNB and monochlorobimane (MCB) without obvious subdivision. However, in panel B, the data points obtained with the alternative substrates are distributed in two main directions. In one of them, the activity with the substrate butylnitrite dominates over those with the substrate MCB, and in the other direction, MCB dominates over butylnitrite.

PRINCIPAL COMPONENT ANALYSIS OF THE SUBSTRATE−ACTIVITY DATA

A principal component analysis (PCA) reveals the main directions of variability of the data set in multidimensional

space (*19*). Principal component 1 (PC1) indicates the direction of the highest variability, and PC2 is the direction of the next-to-highest variability in a direction orthogonal to PC1. The succeeding additional PCs are sequentially orthogonal to all previous directions and mark, one after the other, diminishing variability in the data set. The directions of the principal components are defined by the contributions of the independent variables, i.e., the activities with the alternative substrates. A loading plot demonstrates how the different substrates contribute to the distribution of the data points in relation to the PCs. In other words, the original axes defining the cloud of points in *n*-dimensional space are rotated such that the new PC axes indicate the directions of variability in decreasing order of magnitude, PC1, PC2, PC3, etc. The geometrical interpretation of the loadings is that they express the orientation of the new PC axes with respect to the original axes in *n*-dimensional space. Algebraically, the loadings inform how the original variables are linearly combined to form the PC scores.

Figure 5 illustrates the PCA (A) of the GST data as well as the corresponding loading plot (B) in the first three dimensions. The three corresponding two-dimensional projections are shown below (C−H), to facilitate the interpretation. Three main distributions of the GST variants can be identified (left). In this case, the three clusters are located, as a first approximation, close to the three PC1−PC3 axes, but this is not a general rule. Three of the six parental GSTs are clearly associated with the emerging distributions: hGSTA2 in cluster 2 close to the PC2 axis and hGSTA3 and, to a lesser degree, hGSTA1 in cluster 3 close to the PC3 axis. Cluster 1, near the PC1 axis, contains rGSTA2, but that is less obvious. The two remaining parental GSTs (bGSTA1 and rGSTA3) are found near the center of the data points, where enzyme variants without major distinguishing activities are located. The loading plot (right panel) demonstrates that cluster 1 is characterized by high activities with CDNB, MCB (monochlorobimane), pNPA (*p*-nitrophenylacetate), and PEITC (phenethyl isothiocyanate); cluster 2 with Aza (azathioprine), butylnitrite, and CuOOH (cumene hydroperoxide); and cluster 3 with AD (androsten-3,17-dione). The first three PCs account for 79.6% of the variance in the data set, and components of higher dimensions do not reveal any further clustering of the data.

SCALING OF ACTIVITY DATA

In preparation for PCA, as well as in other forms of multivariate analysis, the data for each variable are commonly normalized to unit variance and centered to a mean of zero. The normalization allows variables measured in small values to have a significant input in the analysis. Otherwise, the data set would be dominated by variables contributing large values. This aspect is particularly important in the search for novel enzyme activities emerging in directed enzyme evolution experiments. The data analyzed in Figures 5−8 have been normalized in this manner.

On the other hand, some investigations may be directed toward the analysis of high-level activities, without regard to minor activities with alternative substrates. In such analyses, normalization to unit variance may not be desirable. Furthermore, analysis of enzyme evolution in natural systems may benefit from scaling activities with respect to the values obtained at ambient concentrations of the alternative substrates. All these scaling procedures relate to the individual columns representing activities with alternative substrates in the data matrix.

NECESSARY AND SUFFICIENT MEMBERS OF THE SUBSTRATE MATRIX

A library of enzyme mutants containing clusters or quasi-species, which can be distinguished by activities with alternative substrates, evidently requires activity measurements with more than one substrate to demonstrate clustering in several dimensions. Identification of quasi-species in *n* dimensions requires at least *n* substrates. On the other hand, *n* substrates could give rise to a higher (or lower) number of clusters. For example, one dimension could demonstrate a trimodal distribution corresponding to high, medium, and low activity. Furthermore, some substrates may give redundant information, such that two (or several) of them could be replaced by one to define the quasi-species.

In this case, Figure 6 shows that three properly chosen substrates (AD, Aza, and CDNB) are sufficient to identify the three quasi-species originally demonstrated with the larger number of substrates (Figure 5). For this identification, it is necessary to select a substrate from each of the three groupings in the loading plot obtained with the complete data set (Figure 5, right). If the substrates are chosen from the same grouping in the loading plot, only one cluster of enzymes will be identified (analysis not shown). Evidently, the recognition of distinct quasi-species is intimately linked to the substrate matrix used in the analysis.

Introduction of activities with a novel substrate, not yet tested, could potentially broaden the substrate−activity space and lead to the emergence of a fourth quasi-species in the data set. However, the resolution in the analysis is limited by the variance caused by the stochastic nature of the experimental data. Judicious experimental design and wisdom in the data analysis are therefore required to maximize the information obtainable from an enzyme library.

SUBSTRATE SELECTIVITIES, EXPRESSION LEVELS, AND FITNESS

In enzyme engineering, as well as in natural evolution, emphasis is often placed on catalytic efficiency in a targeted reaction. A high turnover number of the enzyme-catalyzed reaction is certainly desirable in many instances. However, fitness of an enzyme in a given biochemical system may also be governed by substrate selectivity. Metabolic pathways in a cell, or a chemical process in a bioreactor, can critically depend on the ability of an enzyme to distinguish among alternative substrates. Under these conditions, the importance of substrate specificity usually overrides catalytic efficiency, since low activities with a targeted substrate can be compensated by an elevated enzyme concentration. Figure 7 illustrates how increasing enzyme concentrations enhance catalytic activity with the proper substrate but also, in proportion, the activities with alternative substrates, if they are present.

GABRIEL BIPLOT REPRESENTATION OF VARIANT ENZYMES ACTING ON ALTERNATIVE SUBSTRATES

The data matrix consists of rows representing the different enzyme variants and columns showing the corresponding catalytic activities with the alternative substrates, as described previously. The rows and columns can be regarded as enzyme and substrate vectors, respectively. As a complement to PCA, in which the principal component scores of the enzymes are

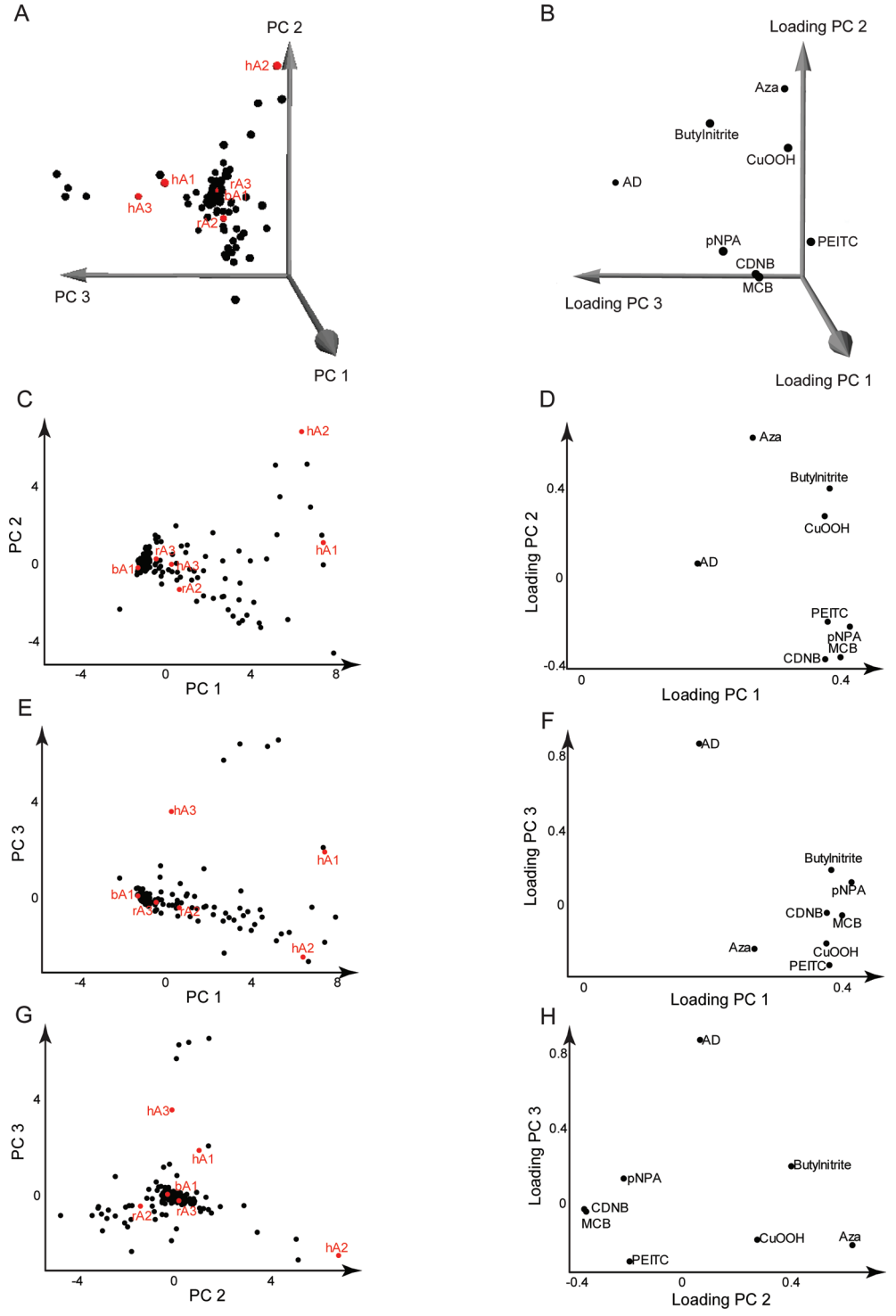

FIGURE 5: Principal component analysis (PCA) of activities of GST variants measured with the eight alternative substrates. The 216 mutants chosen as well as the six parental GSTs were investigated, and the activities were normalized to unit variance and mean centered prior to analysis. The eight-dimensional data are projected as PC scores onto the subspace spanned by the three orthogonal directions, PC1−PC3 (A). The parental GSTs (red dots) are abbreviated as follows: hA1, hGST A1-1; hA2, hGST A2-2; hA3, hGST A3-3; bA1, bGST A1-1; rA2, rGST A2-2; rA3, rGST A3-3. Note the formation of three major distributions, each clustering separately near separate PC axes. The three-dimensional loading plot (B) shows the contributions of the activities with the alternative substrates to the PC scores. For orientation, corresponding two-dimensional plots for each combination of PC1, PC2, and PC3 are shown in panels C, E, and G for PC scores and in panels D, F, and H for PC loadings.

plotted as well as the corresponding loadings (cf. Figure 5), the enzyme vectors and the substrate vectors can be represented simultaneously in one plot. In such a multivariate biplot, the projection is chosen such that the association of enzymes with characteristic substrates is most clearly visualized (*21*). Biplots can be shown in two or three dimensions (Figure 8).

The biplot was designed by Gabriel to represent the rows and columns of a matrix by two sets of vectors in two (or three) dimensions in the same diagram (*21*). When the rank of the matrix is 2 (or 3), the plot is an exact geometrical depiction of the $m \times n$ matrix, but in most cases, the rank will be higher and the corresponding plot based on a rank-2 (or rank-3) "approximation matrix". The approximation is obtained by a singular-value

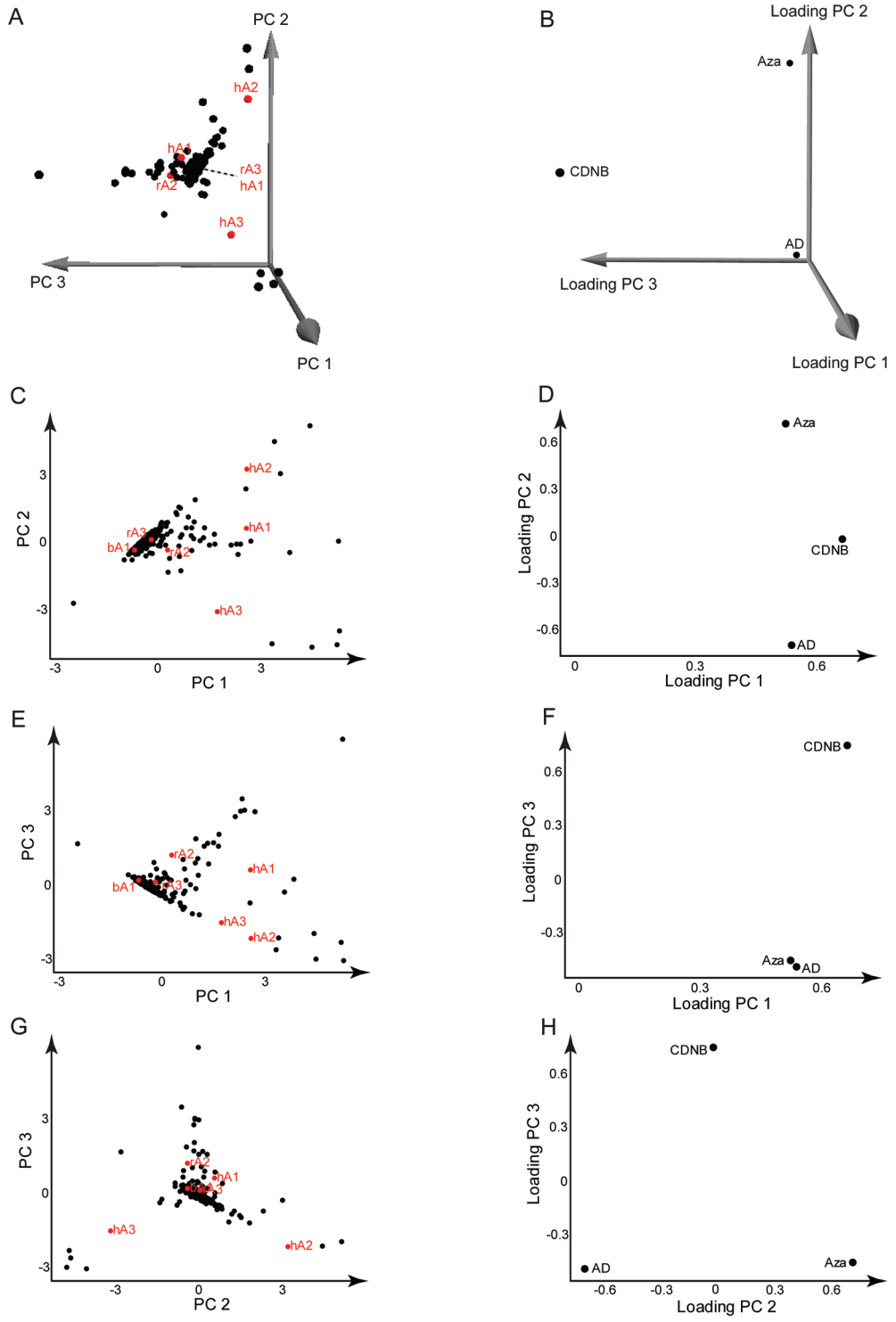

FIGURE 6: PCA of activities of GST variants from a reduced data set based on only three of the eight substrates used. The original data matrix was that used for Figure 5, but only the columns of the three substrates CDNB, Aza, and AD were included in the analysis. The points representing the different GST variants are distributed in three directions, as found with the complete data set in Figure 5. The points are in reverse orientation in relation to the PC1 and the PC3 axes, as a consequence of the altered substrate matrix. However, the quartet beyond hA3 along the PC3 axis in Figure 5 is readily found along the PC1 axis in Figure 6; the singlet along the PC1 axis in Figure 5 is clearly visible along the PC3 axis in Figure 6, whereas the four mutants and hA2 with highest values along the PC2 axis are the same in both figures. For orientation, corresponding two-dimensional plots for each combination of PC1, PC2, and PC3 are shown in panels C, E, and G for PC scores and in panels D, F, and H for PC loadings.

decomposition of the rectangular $m \times n$ matrix such that it is written as the product of two new "row" and "column" matrices, $m \times 2$ and $n \times 2$ (or $m \times 3$ and $n \times 3$ for a three-dimensional plot). Each element in the approximation matrix is the inner product of the row effect and column effect vectors, which can be appraised visually as the product of the length of one vector times the length of the other's projection onto it. The cosine of the angle between arrows approximates the correlation between the variables. Thus, rows (or columns) that are proportional to one another will have the same directions of their corresponding vectors in the biplot. Orthogonal vectors indicate zero correlation between the rows (or columns).

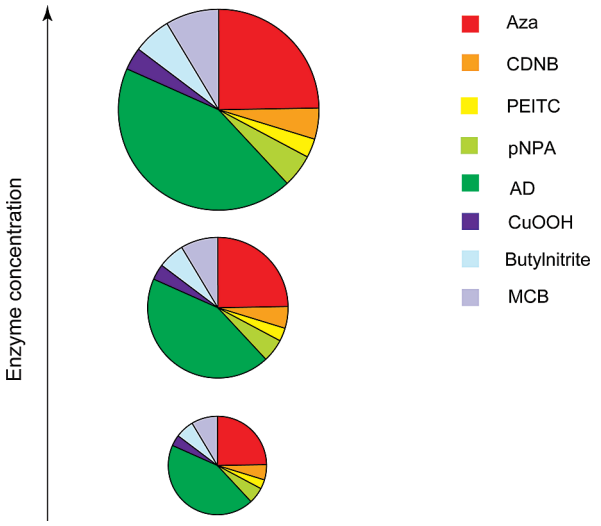

FIGURE 7: Pie chart representation of the substrate−activity profile of a GST variant picked from the library and assayed with eight alternative substrates (cf. Figure 2). The segments in different colors indicate activity of a given substrate as a fraction of the sum of all the specific activities. The area of the chart represents the total catalytic capacity of the GST, and an increased amount of the enzyme will result in a proportionate increase in the catalytic capacity for all alternative substrates.

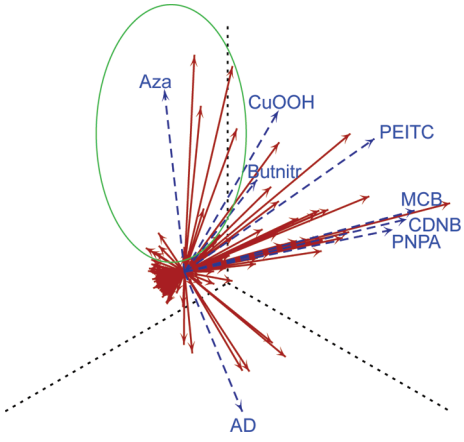

FIGURE 8: Gabriel three-dimensional biplot of multivariate activities obtained with different GST variants in bacterial lysates assayed with eight alternative substrates. The arrows show functional relationships among enzyme variants as well as their associations with the substrates. Quasi-species form bundles of enzyme vectors with similar directions. One of them, characterized by high Aza activity, is indicated by a green ellipse. The data were subjected to unit variance scaling and mean centering, with regard to each substrate separately, prior to the analysis. The lengths of the enzyme vectors (red arrows) are proportional to the amounts of active enzyme present (cf. Figure 7). The substrate vectors (dashed blue arrows) are marked with abbreviations as in Figure 2. In the singular-value decomposition underlying the biplot, the units of the catalytic activity values lose their physical meaning, and the three orthogonal axes are therefore unlabeled. Enzyme vectors show the relative magnitude of the activities and the lengths of substrate vectors the relative loadings of the enzyme vectors.

As in the case of PCA, an adequate representation of the multidimensional vectors in two or three dimensions requires that the major part of the variability of the data be accounted for in these dimensions. Nevertheless, enzyme clusters will be identified as bundles of diverging vectors in the plot, and substrate vectors will indicate associations with the enzyme vectors.

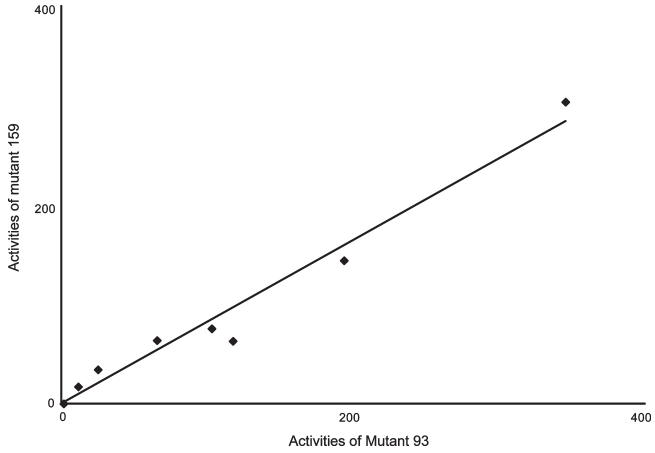

FIGURE 9: Catalytic activities with eight alternative substrates of a crude GST preparation plotted against corresponding activities of another GST preparation. The points indicate values of activities (arbitrary units) in lysates measured with the eight alternative substrates in Figure 2. The enzymes were obtained from clones 159 and 93, which expressed GSTs of identical primary structure. The data fall along a straight line, as expected for identical enzymes, and the slope of 0.83 is a measure of the relative amounts of active GST in the two enzyme samples.

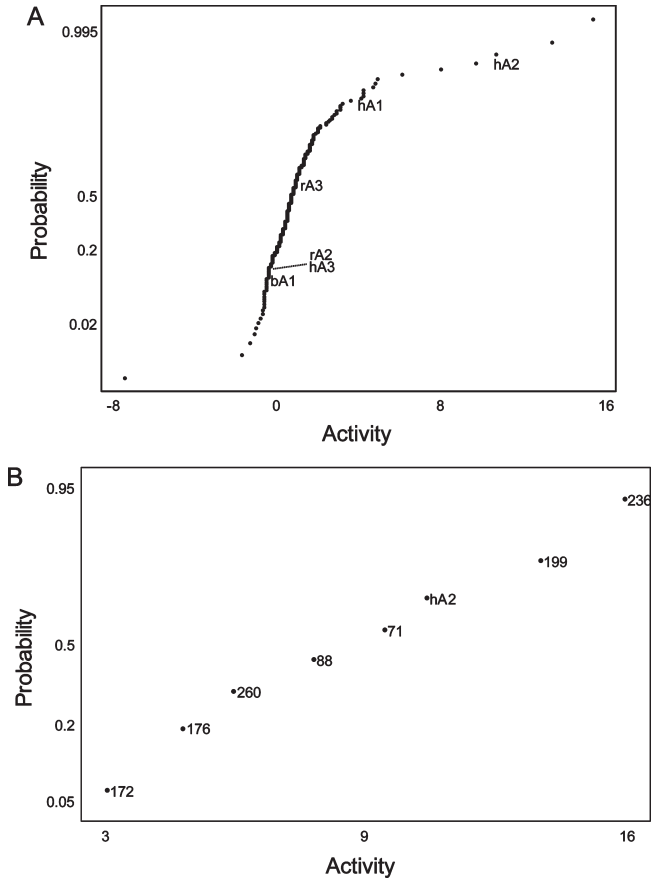

FIGURE 10: Normal distribution probability plots of (A) all 216 data shown in the Gabriel biplot (Figure 8) and (B) the data associated with the quasi-species indicated by an ellipse in the same biplot. The analysis is illustrated in the dimension of the Aza activity. The complete data set (A) clearly does not form a single normal distribution, whereas the quasi-species appears to be Gaussian.

Figure 8 shows a three-dimensional biplot of the enzyme variants randomly chosen from the GST mutant library. Variants with vectors pointing in identical directions represent functionally

identical enzymes, and the length of a vector is proportional to the enzyme concentration. In cases in which the same enzyme has been sampled several times from the library, the activity data will form a cone-shaped bundle of vectors diverging in proportion to the scatter in the experimental data.

In crude biological samples, such as lysates of bacteria expressing recombinant enzyme variants, the concentrations of catalytically active protein may vary due to different expression levels or differences in the efficiency of the recovering active enzyme from the cells. In these cases, the vectors of identical enzyme variants will have the same directions but differ in length in proportion to the concentration of the catalytically active protein. Figure 9 shows a comparison of two crude enzyme preparations obtained from two different bacterial clones expressing the same coding sequence from the GST mutant library. The activities with eight alternative substrates of one of the two preparations are plotted on the *x*-axis, and the activities of the other preparation are plotted on the *y*-axis. A straight line is formed, showing proportionality in the comparison of the diverse activities of the two preparations, as expected, since the expressed GSTs are functionally identical. The slope of the regression line is 0.83, indicating the relative lengths of the vectors, i.e., the ratio of the concentrations of active enzyme in the two samples. In other words, the sample of mutant 159 displays 83% of the catalytic activity of variant 93. For this analysis, the activities have not been normalized, since the relative experimental errors in measurements with substrates giving low activities are higher than in those giving high activities, which would bias the regression analysis. An alternative treatment of the data would be to scale the enzyme vectors to unit length. Functionally identical variants will then have scaled vectors of equal length, and the scaling factors will measure the relative concentrations of the variants.

CHARACTERIZATION OF THE QUASI-SPECIES

The molecular quasi-species was originally introduced in the context of evolving populations of viruses (*22*). It defines the evolving species as a virtual entity typifying a cluster of related variants. A critical principle is that the quasi-species is the true evolving unit, which may be different from the particular variant that shows maximal fitness at a given time point in evolution. In our application of the concept of molecular quasi-species to enzyme evolution, we similarly let the quasi-species represent the evolving cluster of functionally related enzyme variants (*23*). The catalytic properties fluctuate as they evolve in multidimensional activity space, and within limits, functional variants will be regarded as representatives of the same enzyme. However, when the deviations from the standard become too extensive, functional variants have to be considered as novel enzymes.

Evolutionary theory takes into account the assumption that enzyme activities do not have a uniform distribution in functional space but form clusters in a fitness landscape. It is a reasonable proposal that a cluster of enzymes defining a quasi-species could be approximated by a Gaussian multivariate distribution. However, an evolving system will also be characterized by outliers, such that the distribution will turn into a more long-tailed leptokurtic scattering. As indicated above, fitness is defined and probed via the substrate matrix. When the functional properties of the entire population of evolving enzymes are analyzed, their distribution can strongly deviate from a Gaussian multivariate distribution, whereas isolated clusters or quasi-species may show a normal distribution. As an illustration, the normal probability plot in Figure 10A demonstrates the marked deviation from a Gaussian distribution of the activities measured in the samples from the entire GST mutant library. By contrast, the cluster of one of the identified quasi-species (marked with an ellipse in Figure 8) shows a distribution, which is approximately normal (Figure 10B).

SUMMARY AND OUTLOOK

Directed enzyme evolution generally involves mutagenesis accompanied by the search for mutants with valuable properties. This first round is followed by mutagenesis of selected mutants and a new search for desirable enzyme variants. The subsequent recursive process is continued until the targeted goal of evolution has been reached or until no further progress is made. In the breeding of plants and animals, it is well recognized that the genetic background of the progenitors should have a sufficiently broad genetic variation to avoid inbreeding and production of malfunctioning offspring. In an analogous manner, enzyme engineering is also dependent on an optimized variability in the DNA sequences encoding the variant proteins. In other words, it may be important to identify a group of mutants in a given generation of enzyme evolution that can serve as suitable parents for the succeeding generation. A proper identification will optimize the chances of finding mutants with enhanced properties and minimize the risk to entering evolutionary dead ends. The parallel screening with alternative substrates, when feasible, expands the functional landscape and reduces the risk of becoming trapped in local extrema in the optimization procedure.

Our example involving mutants in a GST library illustrates procedures that can be used in the analysis of functional properties of a given generation of enzyme variants. In our laboratory, this approach has been applied for a series of generations and successfully produced GSTs with significantly enhanced activities and altered substrate selectivities (*23*−*30*). It should be noted that additional measured quantities could also be included in the multivariate analysis. They could relate to other functional properties or to physical parameters such as expressivity in cells, binding to other molecular components, stability, etc. Evidently, the outlined multivariate approach is applicable not only to enzymes or other biocatalysts, such as ribozymes or abzymes, but also to the directed evolution of additional entities in biology, chemistry, and physics.

ACKNOWLEDGMENT

Dr. William G. Bardsley (University of Manchester, Manchester, U.K.) has given valuable advice to our development of multivariate analysis of enzyme libraries and provided novel subroutines in SIMFIT (www.simfit.man.ac.uk) for various analyses.

REFERENCES

1. Svendsen, A. (2004) Enzyme Functionality: Design, Engineering, and Screening , Marcel Dekker, Inc., New York.
2. Gerlt, J. A., and Babbitt, P. C. (2009) Enzyme (re)design: Lessons from natural evolution and computation. *Curr. Opin. Chem. Biol. 13*, 10–18.
3. Stemmer, W. P. C. (1994) DNA shuffling by random fragmentation and reassembly: In vitro recombination for molecular evolution. *Proc. Natl. Acad. Sci. U.S.A. 91*, 10747–10751.
4. Tracewell, C. A., and Arnold, F. H. (2009) Directed enzyme evolution: Climbing fitness peaks one amino acid at a time. *Curr. Opin. Chem. Biol. 13*, 3–9.

5. Reetz, M. T., Kahakeaw, D., and Lohmer, R. (2008) Addressing the numbers problem in directed evolution. *ChemBioChem 9*, 1797–1804.
6. Jäckel, C., Kast, P., and Hilvert, D. (2008) Protein design by directed evolution. *Annu. Rev. Biophys. 37*, 153–173.
7. Cho, G. S., and Szostak, J. W. (2006) Directed evolution of ATP binding proteins from a zinc finger domain by using mRNA display. *Chem. Biol. 13*, 139–147.
8. Griffiths, A. D., and Tawfik, D. S. (2006) Miniaturising the laboratory in emulsion droplets. *Trends Biotechnol. 24*, 395–402.
9. Lamboy, J. A., Tam, P. Y., Lee, L. S., Jackson, P. J., Avrantinis, S. K., Lee, H. J., Corn, R. M., and Weiss, G. A. (2008) Chemical and genetic wrappers for improved phage and RNA display. *ChemBioChem 9*, 2846–2852.
10. Crameri, A., Raillard, S. A., Bermudez, E., and Stemmer, W. P. C. (1998) DNA shuffling of a family of genes from diverse species accelerates directed evolution. *Nature 391*, 288–291.
11. Kurtovic, S., Modén, O., Shokeer, A., and Mannervik, B. (2008) Structural determinants of glutathione transferases with azathioprine activity identified by DNA shuffling of alpha class members. *J. Mol. Biol. 375*, 1365–1379.
12. Hansson, L. O., Widersten, M., and Mannervik, B. (1997) Mechanism-based phage display selection of active-site mutants of human glutathione transferase A1-1 catalyzing S_NAr reactions. *Biochemistry 36*, 11252–11260.
13. Demartis, S., Huber, A., Viti, F., Lozzi, L., Giovannoni, L., Neri, P., Winter, G., and Neri, D. (1999) A strategy for the isolation of catalytic activities from repertoires of enzymes displayed on phage. *J. Mol. Biol. 286*, 617–633.
14. Khersonsky, O., Roodveldt, C., and Tawfik, D. S. (2006) Enzyme promiscuity: Evolutionary and mechanistic aspects. *Curr. Opin. Chem. Biol. 10*, 498–508.
15. Josephy, P. D., and Mannervik, B. (2006) Molecular Toxicology, Oxford University Press, Inc., New York.
16. Fersht, A. (1999) Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding , W. H. Freeman & Co., New York.
17. Jensen, R. A. (1976) Enzyme recruitment in evolution of new function. *Annu. Rev. Microbiol. 30*, 409–425.
18. Nath, A., and Atkins, W. M. (2008) A quantitative index of substrate promiscuity. *Biochemistry 47*, 157–166.
19. Krzanowski, W. J. (2000) Principles of Multivariate Analysis , Oxford University Press, Inc., New York.
20. Eigen, M., McCaskill, J., and Schuster, P. (1988) Molecular quasi-species. *J. Phys. Chem. 92*, 6881–6891.
21. Gabriel, K. R. (1971) Biplot graphic display of matrices with application to principal component analysis. *Biometrika 58*, 453–467.
22. Eigen, M. (1996) On the nature of virus quasispecies. *Trends Microbiol. 4*, 216–218.
23. Emrén, L. O., Kurtovic, S., Runarsdottir, A., Larsson, A.-K., and Mannervik, B. (2006) Functionally diverging molecular quasi-species evolve by crossing two enzymes. *Proc. Natl. Acad. Sci. U.S.A. 103*, 10866–10870.
24. Hansson, L. O., Bolton-Grob, R., Massoud, T., and Mannervik, B. (1999) Evolution of differential substrate specificities in mu class glutathione transferases probed by DNA shuffling. *J. Mol. Biol. 287*, 265–276.
25. Broo, K., Larsson, A.-K., Jemth, P., and Mannervik, B. (2002) An ensemble of theta class glutathione transferases with novel catalytic properties generated by stochastic recombination of fragments of two mammalian enzymes. *J. Mol. Biol. 318*, 59–70.
26. Larsson, A.-K., Emrén, L. O., Bardsley, W. G., and Mannervik, B. (2004) Directed enzyme evolution guided by multidimensional analysis of substrate-activity space. *Protein Eng., Des. Sel. 17*, 49–55.
27. Kurtovic, S., Runarsdottir, A., Emrén, L. O., Larsson, A.-K., and Mannervik, B. (2007) Multivariate-activity mining for molecular quasi-species in a glutathione transferase mutant library. *Protein Eng., Des. Sel. 20*, 243–256.
28. Kurtovic, S., Shokeer, A., and Mannervik, B. (2008) Diverging catalytic capacities and selectivity profiles with haloalkane substrates of chimeric alpha class glutathione transferases. *Protein Eng., Des. Sel. 21*, 329–341.
29. Kurtovic, S., Shokeer, A., and Mannervik, B. (2008) Emergence of novel enzyme quasi-species depends on the substrate matrix. *J. Mol. Biol. 382*, 136–153.
30. Mannervik, B., Runarsdottir, A., and Kurtovic, S. (2009) Multi-substrate-activity space and quasi-species in enzyme evolution: Ohno's dilemma, promiscuity, and functional orthogonality. *Biochem. Soc. Trans. 37*, 740–744.