Biocatalytic analysis of biomarkers for forensic identification of ethnicity between Caucasian and African American groups

Friederike Kramer,[a,c] Lenka Halámková,[a,b,#] Arshak Poghossian,[c,d] Michael J. Schöning,[c,d] Evgeny Katz*[a] and Jan Halámek*[a,#]

[a] Department of Chemistry and Biomolecular Science, [b] Department of Biology, Clarkson University, Potsdam, NY 13699, USA. Fax: 1-315-268-6610; Tel: 1-315-268-4421;
E-mails: ekatz@clarkson.edu (EK); jhalamek@clarkson.edu (JH)

[c] Institute of Nano- and Biotechnologies, Aachen University of Applied Sciences, Campus Jülich, Heinrich-Mußmann-Str. 1, D-52428 Jülich, Germany

[d] Peter Grünberg Institute (PGI-8), Research Centre Jülich GmbH, D-52425 Jülich, Germany

[#] New affiliation from September 1st 2013: Department of Chemistry, University at Albany, State University of New York, 1400 Washington Avenue, Albany, NY 12222; email: jhalamek@albany.edu

Electronic supplementary information (ESI)

EXPERIMENTAL SECTION

Chemicals and reagents used:

The following enzymes and organic/inorganic chemicals were purchased from Sigma-Aldrich and used as supplied: creatine kinase from rabbit muscle, Type I (CK, E.C. 2.7.3.2), pyruvate kinase from rabbit muscle, Type III (PK, E.C. 2.7.1.40), L-lactate dehydrogenase from bovine muscle, Type X (LDH, E.C. 1.1.1.27), hexokinase from *Saccharomyces cerevisiae*, Type F-300 (HK, E.C. 2.7.1.1), glucose-6-phosphate dehydrogenase from *Leuconostoc mesenteroides* (G6PDH, E.C. 1.1.1.49), albumin from bovine serum (BSA), serum from a human male (type AB), creatine anhydrous (Crt), phosphocreatine disodium salt hydrate (Crt-P), phospho(enol)pyruvic acid monopotassium salt (PEP), β-nicotinamide adenine dinucleotide phosphate disodium salt ($NADP^+$), β-nicotinamide adenine dinucleotide reduced dipotassium salt (NADH), adenosine 5′-triphosphate disodium salt hydrate (ATP), adenosine 5′-diphosphate

sodium salt (ADP), D-(+)-glucose (Glc), glycyl-glycine and magnesium acetate tetrahydrate. Human serum samples with specific ethnic origin, Caucasian (CA) and African American (AA), were obtained from ProMedDx Specimen Bank (Norton, MA, USA; http://www.promeddx.com) and used according to the ethical regulations established by the Institutional Review Board (IRB). Water used in all of the experiments was ultra pure (18.2 MΩ·cm) from a NANOpure Diamond (Barnstead) source.

Instrumentation and measurements:

A Shimadzu UV-2450 UV-Vis spectrophotometer, with a TCC-240A temperature-controlled holder and 1 mL poly(methyl methacrylate) (PMMA) cuvettes, was used for all optical measurements. The signal corresponding to the concentration of NADH or NADPH was measured optically at $\lambda = 340$ nm.

Composition and operation of the model systems:

(A) System for the two-biomarker CK/LDH-assay (see Scheme 1A in the paper): The system responding to the variable concentrations of CK and LDH was designed and optimized in the present study and realized in 50 mM glycyl-glycine buffer, pH 7.5, containing 2 U/mL PK, 15 mM Crt, 0.6 mM PEP, 0.1 mM NADH and 1 mM ATP. The biocatalytic cascade activated by the CK and LDH inputs resulted in oxidation of NADH to yield NAD^+, thus resulting in the decrease of the absorbance detected optically at $\lambda = 340$ nm. The reaction and optical measurements were performed at 37˚± 0.1 °C.

(B) System for the single biomarker CK-assay (see Scheme 1B in the paper): The system responding to the variable concentrations of CK was realized according to the standard procedure[1-3] in 50 mM glycyl-glycine buffer, pH 7.4, containing 10 U/mL HK and 0.33 U/mL G6PDH, 0.02% (w/v) BSA, 13 mM Crt-P, 1.3 mM ADP, 33 mM Glc, 0.4 mM $NADP^+$ and 4.0 mM magnesium acetate. The biocatalytic cascade activated by the CK input resulted in reduction of $NADP^+$ to yield NADPH, thus resulting in the increase of the absorbance detected optically at $\lambda = 340$ nm. The reaction and optical measurements were performed at 30˚± 0.1 °C. Note that we didn't attempt to redesign/reoptimize the system, using the standard protocol recommended by Sigma-Aldrich[4] and routinely used in hospitals.[5]

Software for statistical computing:

Standard R-project software (free software, version R 2.15.2)[6,7] was used to generate the biomarker (CK and LDH) concentrations mimicking their distribution in CA and AA groups and to analyze the obtained outputs produced in the standard single-biomarker (CK) and new two-biomarker (CK and LDH) assays.

Additional information on the statistical analysis of the experimental data obtained with the model samples:

The histograms (see Figures 3 and 4 in the paper) were characterized by the probability density function (PDF).[8] Since all the distributions are skewed we favored the non-parametric approach over choosing an underlying distribution. The kernel density estimation, which is the most common non-parametric method,[9,10] has been applied. The obtained data were used to generate a PDF curve; this curve depends on the density pattern of the output data. Kernel density estimates were evaluated at an equally spaced grid of 512 points and the code, provided by the R-software,[6] uses fast Fourier transform to convolve the approximation and to evaluate the density fit on the grid. Similarly to the histograms, the bandwidth is a key parameter and the bandwidth's size chosen for the kernel density estimation determines the smoothness of the estimated density. We assessed the impact of different bandwidths on the fitted density holding the kernel fixed. The bias varies markedly with the choice of bandwidth. Because the aim of this study is not to investigate statistical methods we have chosen automatic bandwidths implemented in the R-software that determine the trade-off between the bias and variance and we will not dwell on the details here. In order to mathematically define the quantitative degree of separability of the two probability distributions over the full set of values metric dissimilarity/similarity measures have been calculated. The Hellinger distance (HD) was used to quantify the dissimilarity between two probability distributions and the Bhattacharyya coefficient (BC) was applied as a measure of the similarity between two probability distributions.[11] While various measures have been proposed to compare two histograms,[12] HD and BC used in the present work outperform the others when they deal with a non-Gaussian case.[13] In spite of a strong theoretical background, there is not a given single form used on how to calculate these measures. The approach with the HD/BC calculation can come from a simple parametric model (e.g. like Gaussian), but the underlying distributions of our resulting histograms vary a lot and are rather log-normal. Another option for

the measure comes from the PDF (vector or probabilistic measures).[14] The weakness of the kernel density function (that we used in our histograms to represent the data pattern) originates from the fact that the function is reliable only around the mean/median values and not around extreme values, especially if the sample is small.[15] Based on the nature of our data, we have decided that the histograms would provide the basis for an empirical estimate of the PDFs. The HD/BC have been modified by replacing the integrals with sums to deal with discrete distributions.[16,17] Then, the BC/HD have been used to compare the similarity/dissimilarity between the probability histograms (see Figures 3 and 4 in the paper), where each histogram represents a set of densities of the analytical signal interval (bin). The key step for data processing was normalization of the densities to the sum of all probabilities to be 1. Then, only couples of the bins representing the same signal intervals are compared to form a distance between two considered probability distributions. First, the Bhattacharyya coefficient (BC) determining the relative closeness of the two histograms has been calculated.[14,17,18] Then, the Hellinger distance (HD) has been derived from the BC.[19] Unlike the histograms, where the bin width was made through interactive inspection, for our statistic we preferred to apply a fully automatic procedure which specifies the number and width of the output signal intervals (bins) for both samples (CA and AA group). First, the Freedman and Diaconis (FD) rule, which uses the interquartile range, has been used to produce the interval number and width. The FD approach seems to be a good option if the data are drawn from heavy-tailed distributions.[20,21] Because this situation does not apply to the data obtained from a single-biomarker (CK) assay, where the data are rather spread across the range of the distribution (see Figure 4A in the paper), the Scott's rule has been applied in this case.[22]

Serum stains analysis:

Serum from a human male (type AB; Sigma-Aldrich) was used as supplied to mimic the CA group, while for mimicking the AA group additional amounts of CK (485 mU/mL) and LDH (15 mU/mL) were dissolved in the serum. Therefore, the absolute values of the CK and LDH concentrations were not exactly the same as expected for the CA and AA groups but they represented the difference in the mean values of the CK and LDH concentrations in both groups. The serum samples were dried/aged on a glass surface at 35˚C under reduced air pressure (ca. 25 mm Hg) using a vacuum pump Buchi Vac V500 (Buchi Labs, AG, Switzerland) for various time

intervals (maximum up to 24 hours). Then, the samples were re-dissolved in 0.55 mL of 50 mM glycyl-glycine buffer, pH 7.5, containing 2 U/mL PK, 15 mM Crt, 0.6 mM PEP, 0.1 mM NADH and 1 mM ATP. The obtained solutions were analyzed according to the CK/LDH-assay procedure and the obtained absorbance changes were normalized to the maximum value characteristic for the fresh sample (prior to its drying) mimicking the AA group. The experiment was repeated 5 times for each ageing time period.

Analysis of the human serum samples with the specific ethnic origin:

Human serum samples with specific ethnic origin, Caucasian (CA, 14 samples) and African American (AA, 14 samples), were obtained from ProMedDx Specimen Bank (Norton, MA, USA; http://www.promeddx.com) and used according to the ethical regulations established by the Institutional Review Board (IRB). Pure human serum samples (0.4 mL) were diluted to the ratio of 1:1 with 0.4 mL of 50 mM glycyl-glycine buffer, pH 7.5, containing 2 U/mL PK, 15 mM Crt, 0.6 mM PEP, 1 mM ATP and 0.1 mM NADH. Immediately following the mixing, optical absorbance measurements were recorded continuously at λ = 340 nm, monitoring the decreasing concentration of NADH. Optical measurements were performed at 37.0±0.2 ˚C keeping physiological conditions, and all reagents were pre-incubated at this temperature prior to the measurements.

References

1. S. B. Rosalki and J. H. Wilkinson, *Nature*, 1960, **188**, 1110–1111.
2. L. Noda, T. Nihet and M. F. Morale, *J. Biol. Chem.*, 1960, **235**, 2830–2834.
3. G. Forster, E. Bernt and H. U. Bergmeyer, in: *Methods of Enzymatic Analysis*, (H. U. Bergmeyer, Ed.), 2nd ed., 1974, Vol. II, 789–793, Academic Press, Inc., NY
4. http://www.sigmaaldrich.com/etc/medialib/docs/Sigma/General_Information/creatine_phosphokinase.Par.0001.File.tmp/creatine_phosphokinase.pdf
5. A. H. B. Wu, *Tietz Clinical Guide to Laboratory Tests*, 4th ed., 2006, Saunders-Elsevier, St. Louis, MO.
6. R Core Team, *R: A language and environment for statistical computing*, 2012, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
7. The R Project for Statistical Computing: http://www.R-project.org

8. C. Walck, *Handbook on Statistical Distributions for Experimentalists*, University of Stockholm Internal Report SUF-PFY/96-01 (http://www.stat.rice.edu/~dobelman/textfiles/DistributionsHandbook.pdf).
9. M. Rosenblatt, *Ann. Mathematical Statistics*, 1956, **27**, 832–837.
10. E. Parzen, *Ann. Mathematical Statistics*, 1962, **33**, 1065–1076.
11. A. Basu, H. Shioya and C. Park, *Statistical Inference: The Minimum Distance Approach*, CRC Press, Boca Raton, 2011 (ISBN: 978-1-4200-9965-2).
12. M. M. Deza and E. Deza, *Dictionary of Distances*, Springer, Berlin / Heidelberg, 2009 (ISBN 978-3-642-00233-5).
13. M. J. Wilder, *Automatic target recognition: statistical feature selection of non-gaussian distributed target classes* (Theses), Monterey, California. Naval Postgraduate School, 2010; http://hdl.handle.net/10945/5632
14. S. Cha, *Int. J. Math. Models Methods Appl. Sci.*, 2007, **1**, 300–307.
15. C. Archambeau, *Probabilistic models in noisy environments and their application to a visual prosthesis for the blind*, Ph.D. dissertation, Université Catholique de Louvain, 2005; http://www0.cs.ucl.ac.uk/staff/c.archambeau/publ/phd_ca05.pdf
16. I . Vajda and E. C. van der Meulen, in: Z. A. Karian and E. J. Dudewicz, *Handbook of Fitting Statistical Distributions with R*, Chapman and Hall/CRC, Boca Raton, 2010, pp. 917–994 (ISBN 978-1584887119).
17. K. A. Lee, C. You, H. Li, T. Kinnunen and K. C. Sim, *IEEE Trans. Audio, Speech, and Language Processing*, 2011, **19**, 861–870.
18. F. J. Aherne, N. A. Thacker and P. I. Rockett, *Kybernetika*, 1998, **34**, 363–368 (http://dml.cz/dmlcz/135216).
19. Y. Subaşi and M. Demirekler, Quantitative Measure of Observability for Stochastic Systems, Preprints of the 18th IFAC World Congress, Milano (Italy) August 28 - September 2, 2011, pp. 4244–4249 (available on-line: http://www.nt.ntnu.no/users/skoge/prost/proceedings/ifac11-proceedings/data/html/papers/0829.pdf)
20. D. Freedman and P. Diaconis, *Zeitschrift fur Wahrscheinlichkeitstheorie und verwandte Gebiete*, 1981, **57**, 453–476.

21. Y. Chalabi, D. J. Scott and D. Wuertz, Flexible Distribution Modeling with the Generalized Lambda Distribution, Munich Personal RePEc Archive, MPRA Paper No. 43333, 2012 (http://mpra.ub.uni-muenchen.de/43333/).

22 D. W. Scott, *Biometrika,* 1979, **66**, 605–610.