Bowman, J Pharmacogenom Pharmacoproteomics 2013, 4:1
DOI: 10.4172/2153-0645.1000e134

Journal of Pharmacogenomics & Pharmacoproteomics

Editorial Open Access

Discovering Pharmacogenetic Latent Structure Features Using Divergences

Clive E Bowman*

School of Mathematical Sciences, The University of Reading, Whiteknights, Reading, UK

Abstract

Non-linear divergence functions, as sufficient additive contrast-based measures of direct evidence, offer a smooth universal information basis to deconstruct the stochastic question actually being asked in pharmacogenetic experiments. The orthogonal decomposition of individualized marginal divergences is introduced using entropy and commonalities with PLS-DA. Feature selection is shown in examples of the genetic discriminant analysis of:-up to 3-class diseases; gene by drug treatment studies; and drug-induced multiple adverse events. Analysis over multiple data types, aggregates and dummy indicators is presented. Interaction and epistasis analysis is exemplied. Signal stability, smoothing, approximations and permutation based significance tests are discussed.

Keywords: Log likelihood ratio; SVD; Eigen analysis; Log Bayes factor

The premise in searching for major biological latent structure is that, in Nature (Figure 1), there is truly only a modest set of different types of systems behaviors that will be important to the question at hand. The pattern of these behaviours in humans within and between (sub) systems and measured phenotypes can be parsimoniously decomposed by projection. If substantive, echoes of their impact ought to be found across evolutionary time, and at various levels of human physiology, including disease-comorbidities. Innovative multi-dimensional biomedical computation is needed to dissect such in diseases and drug-response studies. The singular value decomposition (SVD i.e. eigen analysis) has been the mainstay of data projective methods for many years-yet is only patchily deployed in biology and medicine due to its perceived highly technical opaqueness and specialist interpretation. This is a critical barrier to progress in many clinical fields where they are locked into inefficient uni-dimensional experimentation. Moreover, the metric up until now over which people operate SVDs, has been mainly subjective.

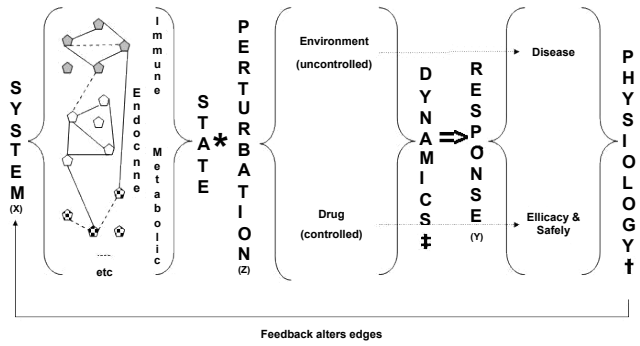

Figure 1: Nature is not static. Left: Genetics drives variation in disease and pharmacogenetic response. The human body comprises hidden (or latent) linked networks of nodal sub-systems with edges representing physiome space (X). Centre: People have different genetically determined system dynamics (response Y_t to a driving function of uncontrolled or controlled perturbations Z), that depend upon the system's physiological state (†). Right: Dynamic response can be measured by phenotypes like: stiff, slow relaxation time, delayed, etc. whether in the domain of disease progression or the development of efficacy or (lack of) safety. Genetic associations (i.e. the network nodes and their interrelations) discovered in an experiment will depend upon:-the choice of (sub)systems (design of X); the choice of phenotypes (design of Y); the partialing out of the confounding kinetics of (or exposure to) the Z drivers (‡) (orthogonalisation and balance of X and Y with respect to Z); and, the (current) system state (i.e. the prevailing network edges) (a priori model for covariance structure constraints for X). *=convolve time varying functions.

Building upon the long forgotten work of Jardine and Sibson [1], Delrieu and Bowman [2] introduced the non-linear mapping of pharmacogenetic data (arising from the exponential family of distributions), into a universal additive information space using divergences. This supervised 'folding' process maps pharmacogenetic or pharmacoproteomic data directly to the question, or contrast of interest posed by the researcher. It shares a philosophical base with maximum likelihood and fold ratio methods. It offers ease of use and direct interpretation of results. It is an integrated analytical frame of reference. Use in the exploratory SVD of high dimensional genetic investigations yields self-correlated sets of biological relevance [3], and results that can be displayed as networks of clear interpretation [4]. Permutation offers a route for empirical significance tests [5].

Deployment in gene × treatment experimentation has already given physiologically interpretable factors [6]. This entropy-based [7-10] transformation allows the simultaneous analysis of multiple phenotypes [11], phenotype severity, and multiple data types/scales [12]. It can be extended to a priori ontologies (aggregates, functional collapsing or blockings), and interaction terms [6]. It highlights and characterizes sample heterogeneity. Partial smoothing and reduced dimensionality approximations are possible [13], and use in genetic epistasis has been described [14]. Recent publications have validated its use in the investigation of multi-phenotype multi-haplotype drug-induced disorders [11,15-17]. It is stable. Dummy variables allow its intuitive dissection. It is topical, state of the art and potentially useful across the breadth of chemical, biological and medical experimentation, as well as clinical, epidemiological, social and economic studies. Easy-to-use

***Corresponding author:** Clive E Bowman, School of Mathematical Sciences, The University of Reading, Whiteknights, Reading, RG6 6AH, UK, E-mail: c.e.bowman@reading.ac.uk

Received June 13, 2013; **Accepted** June 14, 2013; **Published** June 20, 2013

Citation: Bowman CE (2013) Discovering Pharmacogenetic Latent Structure Features Using Divergences. J Pharmacogenom Pharmacoproteomics 4: e134. doi:10.4172/2153-0645.1000e134

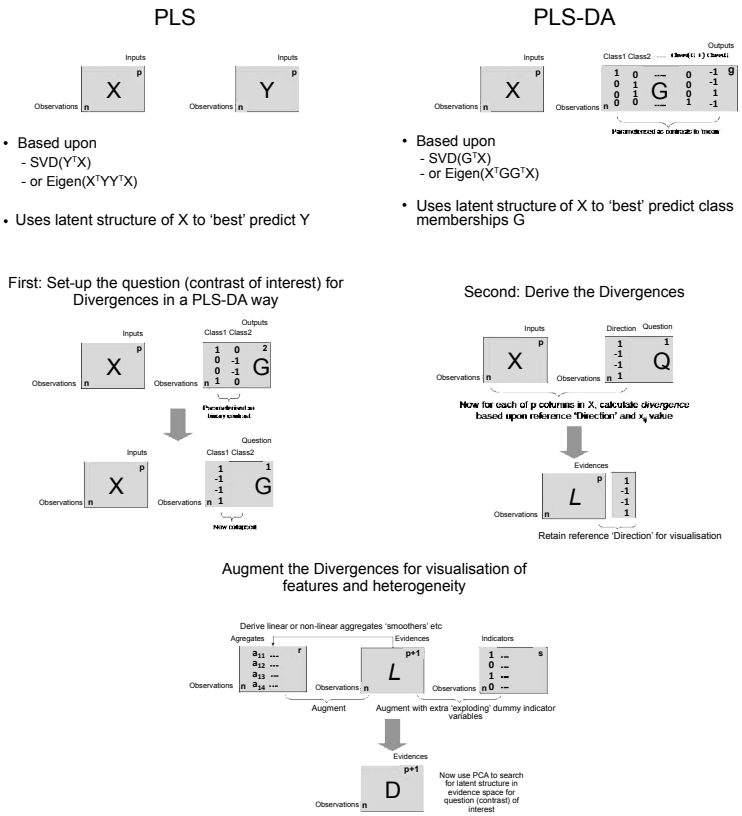

Figure 2: Comparison of divergences method to PLS methods. Upper: PLS variance minimising multivariate projection methods. Left: Projection to Latent Structure (PLS). Right: Projection to Latent Structure discriminant analysis (PLS-DA). Middle and Lower: Block-diagram description of divergences analysis. Middle Left: Setup the research question as a divergence [2]. Middle Right: Derive divergences. For more on the casecont direction

direction [2]. Lower: Augment [6,12] with aggregates, smoothers [2].

smoothers and conditional indicators/questions.

Java-based software for small studies is available from mailto:olivier.delrieu@pgxis.com which is scalable.

Figure 2 Upper illustrates PLS multivariate methods-these have been used to answer comparative research questions-in each case the projection is in the original data measurement space. This has the disadvantage that data variable types should have similar statistical distributions (up to location and scale differences), and that the latent structures found have to be interpreted heuristically to answer the experiment's objectives. A space of uniform properties of direct interpretable relevance to the research question posed is offered by the non-linear transformation based on divergences. This is the innovative idea of a transformation (based upon the entropy of the data space itself) of the original data into an information space of 'the evidence that each observation gives as to a distinction of interest', with the conservation of co-occurrence structure [5]. It's commonality with the philosophy of PLS is clear: PLS depends upon a cross-product of Outputs and Inputs, whilst Divergences fold the Output into the Inputs. The mapping of divergences analysis to PLS-DA is given in Figure 2 Middle and Lower. Unlike PLS, divergences allows the objective simultaneous use of many different data types in a single analysis; given similar error scales then the factor analysis is equivalent to simple PCA (Figure 3).

This data manipulation and evidence based-factor analysis methodology is of use across all biological, clinical and medical domains in recasting many heuristic approaches onto a sound stochastic footing. Inter-operability across scientific experimental disciplines would be ensured by adopting this universal scale. Deployment across the pursuit of biological knowledge and the investigation of the behaviour of living systems would shift current research and clinical practice, and have a measurable direct impact in the quality of our knowledge to extend healthy life and reduce burdens of illness and disability. Clinical adoption would widen the franchise of multidimensional analysis away from highly technical statisticians. A bench mark of success would be a continuing rise of its use, and the validation of its chemical, genetic, proteomic, cellular...organismal. etc. predictions by follow-up laboratory experiments and clinical studies. However, Nature is not static-biological data is different from physical and chemical data-one quintessially has to avoid the 'fundamental error of attribution'. Systems behaviors do not depend just upon their complement of components. Figure 1 shows causality arises from both genetic structure (nodes), and the current physiome state (edges). Phenotypes are a projection of latent genotype structure dependent upon the current physiological state and driving perturbations (i.e. Y=Z(X) over time t, where Y_{t-1} determines edge weightings in deriving Y_t for the current X and Z). Also 'Life' is multi-scale; stochastic phenomena arising simultaneously at all niveaux (although whether in a fractal self-similar way remains to be elucidated). Focusing on biological components and failing to investigate their interactions and the control of these interactions will ensure research success remains illusionary, no matter what analysis algebra is used.

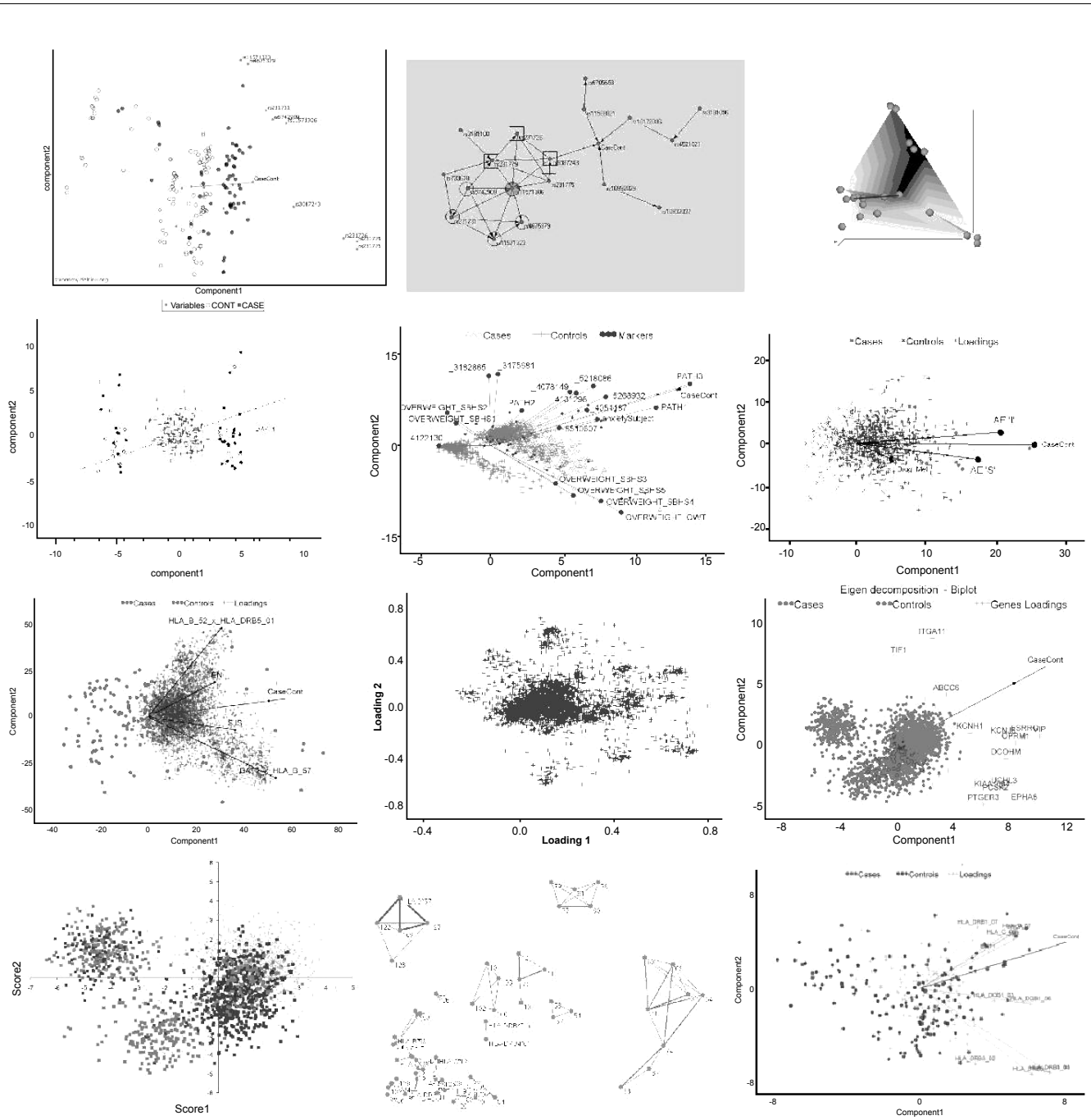

Figure 3: Divergences SVD. Row 1 Left: Biplot of carbamazepine drug-induced HSR. Note upper SNP determinants of severe disease versus SNPs to the right predisposing for mild disease. Middle: Haplotype network of carbamazepine drug-induced HSR showing CTLA4 dominated-versus ICOS dominated-haplotype and one 'lynchpin' SNP in common. Right: 'This-sample' [16,17] permutation tests for significance of SNP effects as a heat map over the ordination, showing only ICOS mutations are of importance to carbamazepine drug-induced HSR. Row 2 Left: Biplot showing drug treated subjects (dark circles) versus placebo controls (open circles), plus importance of DIAPH1 carriage in determining weight loss response to drug. Middle: Simultaneous analysis of SNP, clinical, subject-scored and cellular pathway data across multiple scales in determining psychiatric disease U. Note neurological pathway 3 closely correlated with disease and a subject's own assessment of anxiety. Being overweight has no impact. Right: Simultaneous analysis of two drug-induced adverse events (I versus S) showing that carriage of SNPs in the drug metabolism pathway are related to predisposition to AE 'S'. Row 3 Left: Genetic dissection into two locus haplotypes determining drug induced SJS versus TEN. Note importance of ancestral 57.1 haplotype for SJS, and HLA Class II for TEN predisposition in separating cases and controls [11]. Middle: SVD of divergences interactions over HLA and Alu space in cancer cell line sample showing at least 8 hidden subtypes of causation. Right: Biplot of human disease D showing genetic correlates of disease as well as unexpected control heterogeneity on left. Row 4 Left:

Left: Plot of human disease D controls coloured by clinical centre of subject showing how controls' genetics differs from center to center-confounding investigation of the causes of disease D. Middle: Network of epistatic links in drug induced bullous disorders [11]. Strong links indicate epistatically interacting loci in determining susceptibility. Vertically up the page=predisposition to SJS; Horizontally across the page=predisposition to TEN. Not just haplotype carriage is relevant to the diseases. Right: Simultaneous genetic dissection of multiple drug induced bullous disorders [11].

This paper is self-financed. Some ideas were first presented at PLS09. This field could not have been developed without the efforts of my close friend and collaborator Olivier Delrieu.

References

1. Jardine N, Sibson R (1971) Mathematical Taxonomy. John Wiley and Sons, UK.
2. Delrieu O, Bowman C (2006) Visualizing gene determinants of disease in drug discovery. Pharmacogenomics 7: 311-329.
3. Delrieu O, Bowman CE (2005) Visualisation of gene and pathway determinants of disease In: Quantitative Biology, Shape Analysis, and Wavelets Barber S, Baxter PD, Mardia KV, Walls RE (Ed.) 21-24
4. deNooy W, Mrvar A, Batagelj V (2005) Exploratory social network analysis with Pajek. Cambridge University Press, UK.
5. Bowman CE (2009) Megavariate genetics: What you find is what you go looking for. 19th Altenberg Workshop of Theoretical Biology. Measuring Biology-Quantitative methods: Past and Future. Konrad Lorenz Institute, Austria.
6. Delrieu O, Bowman CE (2006) Visualisation of gene by gene interactions in pharmacogenetics. Poster at International Congress of Human Genetics, Brisbane, Australia.
7. Shannon CE (1948) A mathematical theory of communication. Bell Syst Techn J 27: 379-423
8. Kullback S (1959) Information theory and statistics. Dover Books on Mathematics.
9. Kullback S, Leibler R (1951) On information and sufficiency. Ann Appl Stat 22: 79-86.
10. Tribus M (1961) Thermodynamics and thermostatics: An introduction to energy, information and states of matter, with engineering applications. D. Van Nostrand Company Inc., New York, USA.
11. Pirmohamed M, Arbuckle JB, Bowman CE, Brunner M, Burns DK, et al. (2007) Investigation into the multidimensional genetic basis of drug-induced Stevens-Johnson syndrome and toxic epidermal necrolysis. Pharmacogenomics 8: 1661-1691.
12. Delrieu O, Bowman C (2007) On using the correlations of divergences In: Systems Biology and Statistical Bioinformatics, Barber S, Baxter PD, Mardia KV (Ed.) 27-35.
13. Charalambous C, Delrieu O, Bowman C (2008) Whole genome scan algebra and smoothing. In: The Art and Science of Statistical Bioinformatics Barber S, Baxter PD, Gusanto A, Mardia KV (Ed.) 21-27.
14. Bowman CE, Delrieu O (2009) Correlation laplacians, haplotype networks and residual pharmacogenetics. In: Statistical Tools for Challenges in Bioinformatics. Gusanto A, Mardia KV, Fallaize CJ (Ed.) 25-31.
15. Alfirevic A, Vilar FJ, Alsbou M, Jawaid A, Thomson W, et al. (2009) TNF, LTA, HSPA1L and HLA-DR gene polymorphisms in HIV-positive patients with hypersensitivity to cotrimoxazole. Pharmacogenomics 10: 531-540.
16. Bowman C, Delrieu O (2009) Immunogenetics of drug-induced skin blistering disorders. Part I: Perspective. Pharmacogenomics 10: 601-621.
17. Bowman C, Delrieu O (2009) Immunogenetics of drug-induced skin blistering disorders. Part II: synthesis. Pharmacogenomics 10: 779-816.