Xu *Appl Inform* (2016) 3:4
DOI 10.1186/s40535-016-0020-3

Applied Informatics

RESEARCH

Open Access

Enviro-geno-pheno state approach and state based biomarkers for differentiation, prognosis, subtypes, and staging

Lei Xu[1,2*]

*Correspondence: lxu@cs.sjtu.edu.cn
[1] Department of Computer Science and Engineering, Centre for Brain-inspired Computing and Bio-Health Informatics, The School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, SEIEE Building 3, 800 Dongchuan Road, Minhang District, 200240 Shanghai, China
Full list of author information is available at the end of the article

Abstract

Finding biomarkers for differentiation, prognosis, subtypes, and staging takes a key role in precision medicine, usually featured by association analysis on geno-measures and pheno-measures. Recent efforts turn to identifying the role of a biomarker under certain condition or in a particular environment, represented by a set of enviro-measures. This paper proposes to consider the joint domain of geno-measures, pheno-measures, and enviro-measures, in which one element (i.e., each triple jointly taken by the three measures) represents a possible behaviour of the bio-system under investigation. A collection of elements that locate adjacently and share a common system status represents a 'state', and the system is characterised by a number of such states learned from samples. Instead of directly using one or a set of geno-measures as a biomarker, such an enviro-geno-pheno state (E-GPS) is considered as a biomarker, indicating 'health/normal' versus 'risk/abnormal' together with its associated enviro-geno-pheno conditions. Association analyses for differentiation, prognosis, subtypes, and staging can be performed between such E-GPS biomarkers and those measures representing clinical phenotypes and treatments, made either on one state or cross multiple states. Moreover, potential applications are suggested for analyses of expression data, sequencing data, and their integrative uses.

Keywords: Enviro-geno-pheno state, Biomarker, Differentiation, Prognosis, Stage, Subtype, Integrative study, Case–control study, Genome-scale sequencing, Expression profile analysis

Background

In an extended sense, we use a geno-measure g (shortly g-measure) to refer a genetic measure that takes either a real value or one of a few labels, e.g., the expression level of a gene, the frequency of a mutation, the genotype of an SNP, etc. Moreover, g-measure can also be $\mathbf{g}$ that denotes a vector or a matrix with each element being one of such genetic measures. On the other hand, we use a pheno-measure ϕ (shortly ϕ-measure) to refer a phenomenon indicator that is typically a categorical label or an integer number, indicating different subtypes or stages of a cancer or complex disease. Moreover, we may use a real-valued ϕ for a phenomenon that has a large category size or is directly featured by a continuous measure, e.g., survival length. Considering multiple phenomena jointly,

Springer Open

a ϕ-measure could also be a vector $\boldsymbol{\phi}$ with each element being such an individual phenomenon indicator.

A biomarker that identifies abnormal or normal is a g-measure g that demonstrates a significant difference between the case population and the control population, and a biomarker that indicates subtypes or stages is one g-measure g that demonstrates a significant characteristic underlying samples of each corresponding group, while a biomarker of prognosis provides a good prediction on post-treatment survival. Some is a common biomarker that is useful to all the uses, while some particularly works for merely one or two of them.

Typically, a g-measure g is an SNP in GWAS or an expression value of a gene in expression profile. Moreover, a g-measure can be a vector $\mathbf{g}$ that consists of multiple SNPs in a segment of DNA sequence, where the segment corresponds to a gene or a noncoding RNA (lncRNA, circRNA, etc.) in consideration. In addition, $\mathbf{g}$ may consist of a number of features obtained from mutation analysis. For analysing expression profile with the tumour versus its paired adjacent tissue, $\mathbf{g}$ is a two-dimensional vector that consists of simply the expressions of tumour and of the paired adjacent tissue, see page 36 in Ref. Xu (2015a) and Fig. 7 in Ref. Xu (2016). Even generally, a vector $\mathbf{g}$ may represent a bio-unit in consideration, which covers more than one gene, e.g., expressions of several mRNAs that group in a signature on a heart map or certain features that represent one biological functional module.

Conventionally, whether a g-measure acts as a biomarker was examined on a set of case–control samples. For examples, a gene expression biomarker for prognosis takes a high value to indicate positive to survival. Alternatively, there may also be a biomarker that takes a low value to indicate positive to survival. Traditionally, such a biomarker reflects a difference between samples of the g-measure without particularly taking a specific condition in consideration. Recently, multiple reasons appear to support that it would be better to examine a biomarker under certain conditions or in a particular environment.

First, the meaning of a biomarker identified unconditionally may change considerably in a particular environment. An example is one recent finding of CDX2 as a gene expression biomarker for prognostic in colon cancer (Dalerba et al. 2016). Without considering any condition, its high expression is preferred because the rate of 5-year disease-free survival with stage II CDX2-negative colon cancers was significantly lower than the rate with stage II CDX2-positive colon cancers. Interestingly, it was found that the rate of 5-year disease-free survival with stage II CDX2-negative tumours who were treated with adjuvant chemotherapy became significantly higher than the rate with ones who were not treated with adjuvant chemotherapy, i.e., a low expression became preferred under the condition when adjuvant chemotherapy was treated.

Second, the role of some biomarker that is unable to be unconditionally identified will become detectable under a specific condition. One example is one recent study on IDH1-mutant glioma malignant progression (Bai et al. 2016). Considering all 82 sequenced gliomas conditioning on that they all have IDH1 mutations, the role of rare mutations of NOTCH1 and NOTCH2 was identified, occurring within sequences encoding the EGF-like domains, which is consistent with inactivating mutations identified in squamous cell carcinomas.

Third, a comprehensive study demands jointly considering a biomarker that consists of multiple g-measures or even jointly considering multiple biomarkers, for which one effective way is hierarchical formulation. In its simplest case, jointly considering two parts can be made subsequently by first considering one part and then considering the rest part conditioning on the first part. One example is the recent molecular analysis of gastric cancer that identifies four subtypes of gastric cancer by a binary tree with three layers (Cristescu et al. 2015), where subtypes MSS/TP53+ and MSS/TP53− are identified by an integrated biomarker named TP53 signature conditioning on an integrated biomarker named EMT signature and an integrated biomarker named MSI signature.

In a summary, a geno-phenotype study involves not only g-measures and ϕ-measures, but also a set $\mathbf{e}$ of enviro-measures (shortly e-measures) that specify certain condition or a particular environment underlying the study. In other words, we actually make an enviro-geno-pheno integrative study, which may be shortly denoted by a notation $g \underset{e}{\rightarrow} \phi$ or $\mathbf{g} \underset{\mathbf{e}}{\rightarrow} \boldsymbol{\phi}$, where each e-measure may represent one of treatments, patient characteristics or g-measures jointly in consideration.

In the rest of this paper, we propose a generic approach as summarised in Table 1. First, the approach identifies one or several convex subsets in the joint domain of g-measures, ϕ-measures, and e-measures, with each subset representing a state of the bio-subsystem in our investigation. Shortly, such a state is called enviro-geno-pheno state (E-GPS) that acts as E-GPS biomarker, indicating 'health/normal' versus 'risk/abnormal' together with its associated enviro-geno-pheno conditions. Second, the approach makes association analysis from such E-GPS states to not only ϕ-measures or clinical phenotypes but also e-measures, towards various tasks that include but are not limited to differentiation, prognosis, subtype, and staging.

Even generally, g-measures may not only be limited to genetic measures but could be also other measures that serve as the inner ground of study, called ground measures (still g-measures shortly). In other words, the E-GPS approach is also applicable to those data-mining tasks that can be formulated into the format $g \underset{e}{\rightarrow} \phi$.

Methods

Whether a living system survives healthily or a machine system runs normally is featured by its internal status that could be one of several types. One major type is 'health/good/normal' or negative '−', the other type is 'risk/bad/abnormal' or positive '+' . There could be other types too, e.g., sub-health or slightly abnormal. In addition, there may be a type indicating 'unknown/confusing' or shortly '?'.

Specifically, a system status is measured via a set $\mathbf{g}$ of internal intrinsic or ground factors and a set $\mathbf{e}$ of environmental factors, as well as a set $\boldsymbol{\phi}$ of the external behaviours or phenotypes that the system demonstrates correspondingly. Let $\mathbf{G}, \mathbf{E}$, and $\boldsymbol{\Phi}$ to indicate the domain of $\mathbf{g}$, $\mathbf{e}$ and $\boldsymbol{\phi}$, respectively, as illustrated in Fig. 1a, a system variate $\xi \in \mathcal{D}_{g\phi e} = \mathbf{G} \times \mathbf{E} \times \boldsymbol{\Phi}$ represents an enviro-geno-pheno triple and is associated with a label, e.g., coloured green for 'normal' and coloured red for 'abnormal' as illustrated in Fig. 1b, that indicates an instance of the system status. Moreover, a subset $R_s \subset \mathcal{D}_{g\phi e}$ conceptually describes a possible relation among $\mathbf{g}$, $\mathbf{e}$, and $\boldsymbol{\phi}$. Not all possible subsets are interesting. We are interested in that R_s is convex and every element in R_s shares a same type of system status. The system behaves the same as long as ξ locates within R_s, namely R_s represents one Enviro-Geno-Pheno state (E-GPS) in $\mathcal{D}_{g\phi e}$, shortly denoted by s. The system behaviour is actually an external manifestation of one or several such states, as

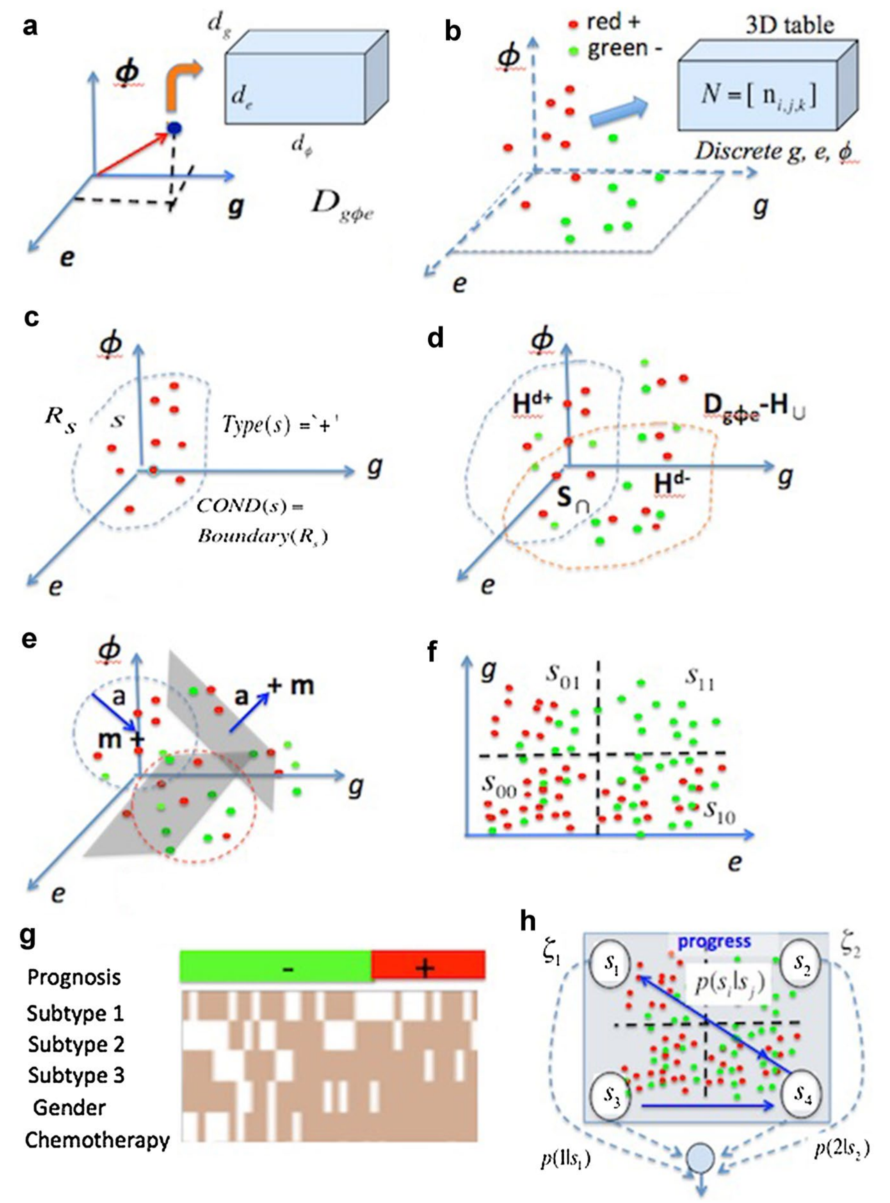


Fig. 1 Enviro-geno-pheno state as biomarker, shortly E-GPS biomarker **a** Each element of $\mathcal{D}_{g\phi e}$ is generally a $d_g \times d_e \times d_\phi$ data cubic, where d_g, d_e, and d_ϕ are the dimensionalities of **g**, **e**, and $\boldsymbol{\phi}$, respectively. **b** When *g*, *e*, and ϕ are univariate, the case is illustrated by a scattering map, which is degenerated into an $m_g \times m_e \times m_\phi$ table that represents a discrete distribution when *g*, *e*, and ϕ take m_g, m_e, and m_ϕ discrete values, respectively. **c** A convex set R_s acts as E-GPS biomarker, with the system status indicated by *Type*(*s*) and the boundary condition by *COND*(*s*) about genotypes, phenotypes, and environments by the boundary of R_s. **d** The possible system statuses are featured by E-GPS states that are learned from given samples, by minimising the criterion given by Eq. (1) or (4). **e** For a finite size of samples, we prefer a simple parametric model, e.g., by one of the two choices given in Eq. (7). **f** An E-GPS state corresponds to a convex subset with all its elements dedicated to the same status type, e.g., s_{11} is a biomarker of '*green*', which maybe relaxed to require a probabilistic dedication, i.e., samples falling in a convex subset are mostly dedicated to a same status type. Contrastingly, a c-state is featured by that two status type compete samples, e.g., s_{01} and s_{10}. **g** Prognosis analysis can be made per d-state, as addressed in Table 1 (3)(a). In addition, subtype analysis is made per state, with the *top row* indicating '*green*' and '*red*' samples and other *rows* indicating subtypes in binary values. The relation between the E-GPS state in consideration and each subtype is examined by their intersection. **h** We may compare the configuration of states jointly. In addition, the results of phenotype analysis per state can be combined, with help of the weighting probability $p(j|s_j)$ in accordance with the individual performance of each state. We may further make state transient analysis by estimating the transfer probabilities $p(s_i|s_j)$

illustrated in Fig. 1h. For each E-GPS state s, not only its associated type $Type(s)$ indicates the system status, for which we subsequently focus on $Type(s)$ from one of values $-, +, ?$ for simplicity, and the study can be rather straightforwardly extended to other sub-health types, but also the boundary of its corresponding convex set R_s describes the condition $\mathcal{B}_s = COND(s) = Boundary(R_s)$ to stay at this state, as illustrated in Fig. 1c.

Requiring that every element in R_s shares a same type of system status, an E-GPS state is featured by its dedication to one specific type of system status, and thus is shortly called a dedicating state or shortly d-state, e.g., the green d-state $s^{(11)}$ in Fig. 1f dedicates to a 'normal' system status. To tolerate some error or disturbance, we may relax to require that every element in R_s gets a high enough probability to share a same type of system status, that is, we consider the concept of d-state in a probabilistic sense, e.g., the red d-state s_{00} in Fig. 1f dedicates to one 'abnormal' system status. In addition to d-states, we may also need to handle subsets confused with different types or unknown types of samples, shortly we also regard such a subset as a confusing state or c-state, e.g., s_{10} in Fig. 1f.

Instead of adopting a standard routine that directly uses one or a set of g-measures as a biomarker of phenotypes that we aim at, we suggest to use each d-state as a biomarker, shortly, called E-GPS biomarker. Its difference from considering merely **g** measures as a biomarker lays in not just jointly considering $\mathbf{g}, \mathbf{e}$ as a biomarker. Even without **e**, an E-GPS biomarker $R_s \subset \mathcal{D}_{g\phi e}$ degenerates to a binary relation or a subset of $\mathbf{G} \times \mathbf{\Phi}$ whilst we traditionally consider a special bi-relation called function $F : \mathbf{G} \rightarrow \mathbf{\Phi}$ or $\boldsymbol{\phi} = f(\mathbf{g})$. Actually, widely studied is a linear or logistic function $f(\cdot)$, which is an example of merely considering **g** measures as a biomarker. In other words, an E-GPS biomarker extends such a function not only to a bi-relation but also further to a triple-relation $R_s \subset \mathcal{D}_{g\phi e}$ with **e** also taken in consideration, featured by the corresponding condition $\mathcal{C}_s$ that summarises the boundary conditions about genotypes and phenotypes as well as environments.

A representation of R_s is learned from a given set of samples, for which we may consider a d-state s to be described by the convex hull of samples of the corresponding type, as illustrated in Fig. 1c. It is better to jointly consider a d-state of type '+' (shortly s^{d+}) by the convex hull $\bar{H}^{d+}$ of red samples and a d-state of type '−' (shortly s^{d-}) by the convex hull $\bar{H}^{d-}$ of green samples, as illustrated in Fig. 1d. There may be a nonempty or rather large intersection $\bar{S}_\cap$ that should be cut away from both $\bar{H}^{d+}$ and $\bar{H}^{d-}$, for which we shrink both $\bar{H}^{d+}$ and $\bar{H}^{d-}$ into a convex subset $H^{d+} \subseteq \bar{H}^{d+}$ and a convex subset $H^{d-} \subseteq \bar{H}^{d-}$ with minimal intersection $S_\cap$ but a maximal union $S_\cup$ such that the red samples and the green samples are best represented by H^{d+} and H^{d-}, respectively, which is implemented by minimising the following criterion

$$J(s^{d+}, s^{d-}) = \frac{|S_\cap|}{|S_\cup|}, \quad S_\cap = H^{d+} \cap H^{d-}, \quad S_\cup = H^{d+} \cup H^{d-}. \tag{1}$$

As a whole, H^{d+} and H^{d-} jointly divide $\mathcal{D}_{g\phi e}$ into four subsets

$$\begin{aligned} &\mathcal{S} = \{S_-^{d+},\ S_-^{d-},\ S_\cap, \neg S_\cup\}, \\ &S_-^{d+} = H^{d+} - S_\cap,\ S_-^{d-} = H^{d-} - S_\cap,\ S_\cap, \quad \text{and} \quad \neg S = \mathcal{D}_{g\phi e} - S_\cup, \end{aligned} \tag{2}$$

which includes those special cases of three subsets when $S_\cap$ becomes one of H^{d+} and H^{d-}, and also those special cases of two subsets when H^{d+} and H^{d-} are identical.

No longer each subset $S \in \mathcal{S}$ is guaranteed to be convex. Without requiring such a convexity, we further consider each subset by the following ratio of minority

$$r_S = \frac{\min\{n_S^+, n_S^-\}}{n_S}, \quad n_S = n_S^+ + n_S^-, \tag{3}$$

where there are a number n_S of the samples in $S \in \mathcal{S}$, with n_S^+ red samples and n_S^- green samples, respectively. We may regard S as a d-state when r_S goes below a threshold γ_s and n_S is bigger than a minimum number n_0. In this case, we expect to minimise r_S for every $S \in \mathcal{S}$. Possibly, there is a subset S with its r_S being impossibly reduced below γ_s. Forcibly minimising such a r_S will unfavourably increase the competing ratios of other subsets. Thus, we are better to leave this r_S away from being minimised. Considering every $S \in \mathcal{S}$ jointly, we minimise the following criterion

$$J(\mathcal{S}) = \sum_{S \in \mathcal{S},\ s.t.\ n_S \geq n_o\ \&\ r_S \leq \gamma_S} \varepsilon(r_S, d_S, \eta_S), \tag{4}$$

where $\varepsilon(u, v, w) \geq 0$ is a function with

$$\frac{\partial \varepsilon}{\partial u} \geq 0, \quad \frac{\partial \varepsilon}{\partial v} \leq 0, \quad \frac{\partial \varepsilon}{\partial w} \leq 0. \tag{5}$$

That is, a smaller value J prefers a smaller r_S or equivalently a d-state. Moreover, d_S reflects a degree of separation between the samples inside and outside S. It follows from $\frac{\partial \varepsilon}{\partial v} \leq 0$ that a smaller value J prefers bigger d_S. Furthermore, η_S reflects a degree of balance on the numbers of samples over subsets in $\mathcal{S}$, e.g., we may consider the following entropy

$$\eta_S = -\frac{1}{\#\mathcal{S}} \sum_S \frac{n_S}{\sum_S n_S} \ln \frac{n_S}{\sum_S n_S} \quad \text{or} \quad \eta_S = -\frac{1}{\#\mathcal{S}} \sum_S \left[\frac{n_S}{\sum_S n_S}\right]^2. \tag{6}$$

It follows from $\frac{\partial \varepsilon}{\partial w} \leq 0$ that a smaller value J prefers bigger η_S or a configuration with a least number of states and also with samples dedicated to the states in a balanced way.

In implementation, there could be different choices for representing H^{d+} and H^{d-}. Typically, a finite size of samples restrains our preference to a simple parametric model. Two simplest choices are given as follows:

(a) a hyper-sphere parameterized by a location vector $\mathbf{m}$ and a radius a, (7)

(b) a hyper-plane parameterized by a location vector $\mathbf{m}$ and a normal vector $\mathbf{a}$.

For examples, Choice (a) is illustrated by the dashed circles in Fig. 1e, and Choice (b) is illustrated by the greyed planes in Fig. 1e. The former choice is similar to the general case in Fig. 1d, parameterised by a least number of free parameters to be estimated by minimising the criterion given by Eqs. (1) or (4). However, there are two limitations. First, the spherical shape is not suitable for representing a sample population in an elongated configuration featured by some orientation. Second, there may be some subset in Eq. (2) that is not convex and thus loses the robustness of a convex set.

Though the first limitation may become broken with hypersphere replaced by hyper-ellipse, not only it largely increases the number of free parameters and thus becomes prone to overfitting but also some subset in Eq. (2) may still not be convex. Favourably, Choice (b) gets a small incremental in free parameters, i.e., simply with the scalar a replaced by a vector $\mathbf{a}$, such that the second limitation is overcome and the first limitation is at least partially overcome. Specifically, $\mathcal{D}_{g\phi e}$ is partitioned into at least two convex subsets and at most four convex subsets by two hyper-planes, and the resulted subsets may also have some orientation. Again, we may estimate the two hyper-planes by minimising the criterion given by Eqs. (1) or (4).

Illustrated in Fig. 1f is a simple example that $\mathcal{D}_{g\phi e}$ is partitioned into four convex subsets S_{11}, S_{01}, S_{10}, and S_{00} by two lines. Specifically, the subset S_{11} represents a d-state s_{11} as good biomarker of 'green' (i.e., normal), though unconditionally using g as a biomarker cannot differentiate the normal versus the abnormal. As a d-state, samples of the state s_{11} are all dedicated to 'green', while the state s_{00} is almost a d-state that corresponds to the subset S_{00} that consists of mostly red samples. The other two subsets S_{01} and S_{10} act as the c-states. Relaxing two lines in Fig. 1f to become adjusted freely, an optimal partition may be obtained by minimising the criterion given by Eq. (4).

This approach differs from the conventional linear discriminating analysis not just in that one hyper-plane is replaced by two hyper-planes, but also in that the classification error is replaced by the dedication degree of samples while the c-states are excluded from disturbing the d-states. Then, the samples of each c-state maybe further divided into two to four convex subsets using this approach too, e.g., the subset S_{01} in Fig. 1f can be further divided into two d-states still within $\mathcal{D}_{g\phi e}$. Recursively doing so, we are led to a tree as illustrated in Fig. 2e. As a whole, red samples are represented jointly by a number of d-states, either in a probabilistic combination as illustrated in Fig. 1h or via a union of convex subsets while this union may be no longer convex. Similarly, red samples are represented by a number of d-states too.

To avoid overfitting, in Eq. (4) we impose a lower bound on the number of samples in each d-state. In addition, we may merge samples of adjacent c-states to form a big c-stage before dividing one c-state into subsets on the next level, as illustrated in Fig. 2f. For a small size of samples, we may further reduce the number of free parameters by restraining two hyper-planes in parallel, i.e., reduce one orientation vector $\mathbf{a}$ into a scalar a to denote the distance between two parallel hyper-planes. Learning may be simplified into a two-stage implementation as illustrated in Fig. 2c, d. First, the normal direction of parallel hyper-planes is learned either directly by support vector machine (SVM) (Suykens and Vandewalle 1999; Suykens et al. 2002) as shown in Fig. 2c or with help of Fisher discriminative analysis (FDA) as shown in Fig. 2d. Second, samples are projected onto the normal direction and further divided into three subsets by minimising a simplified version of $J(S)$ given by Eq. (4), as shown in Fig. 2a, b.

In Eq. (3), one sample within S will contribute one count to either n_S^+ or n_S^- regardless of where it locates. We may consider that each sample $x \in S$ is associated with weight coefficients $w^+(x)$ for red samples and $w^-(x)$ for green samples, based on which we modify n_S^+ and n_S^- in Eq. (3) into the follow ones:

$$\begin{aligned} n_S^+ &= \sum_{\text{each red sample } x\in S} w^+(x), \\ n_S^- &= \sum_{\text{each green sample } x\in S} w^-(x). \end{aligned} \tag{8}$$

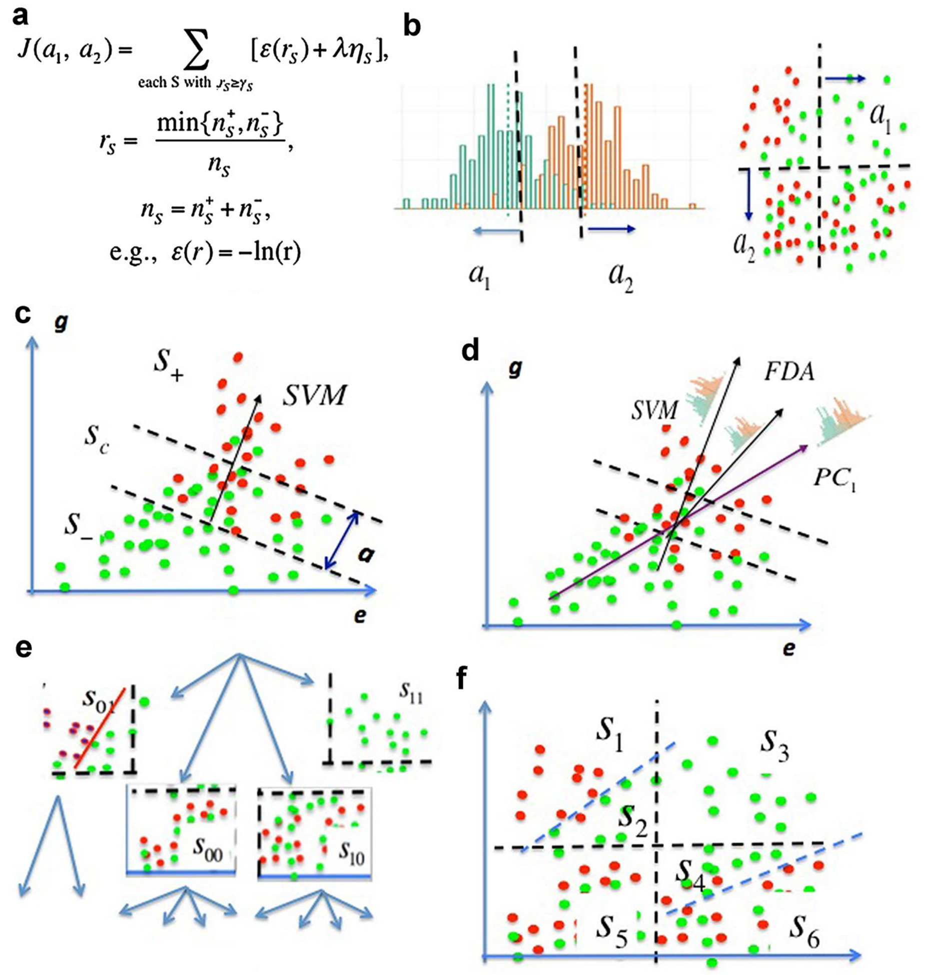

a
$J(a_1, a_2) = \sum_{\text{each S with } r_S \geq \gamma_S} [\varepsilon(r_S) + \lambda \eta_S],$
$r_S = \frac{\min\{n_S^+, n_S^-\}}{n_S},$
$n_S = n_S^+ + n_S^-,$
e.g., $\varepsilon(r) = -\ln(r)$
b
a_1
a_2
c
g
e
S_+
S_c
S_-
SVM
a
d
g
e
SVM
FDA
PC_1
e
s_{01}
s_{11}
s_{00}
s_{10}
f
s_1
s_2
s_3
s_4
s_5
s_6

Fig. 2 Learning E-GPS- and E-GPS-based analysis **a** Learning E-GPS with $J(\mathcal{S})$ given by Eq. (4) simplified into $J(a_1, a_2)$, where η_S is given by Eq. (6) and $\lambda \geq 0$ is a weight for the role of η_S. **b** Two lines in special cases with only two free parameters. **c** We may get three convex subsets by considering two parallel lines featured by three free parameters. One way for learning the two lines is using support vector machine (SVM) (Suykens and Vandewalle 1999; Suykens et al. 2002) for several different choices of the margin a among which we pick the best one according to J. **d** Alternatively, two parallel lines may also be obtained in two steps. First, we find the normal direction of the two lines by SVM and then project all the samples orthogonally onto the direction. Second, we treat the projected samples in the same way as in (**b**). Instead of SVM, the normal direction may also be determined by either Fisher discriminative analysis (FDA) or principal component analysis (PCA). **e** Recursively, we perform the division as illustrated in (**c**) or (**d**) on each c-state, ..., so on so forth, until all the remaining c-states are noncuttable, see Table 1(1)(c). Finally, we get a tree with each d-state (e.g., s_{11}) as a leaf. **f** Defragment is made from time to time by merging adjacent d-states s_2, s_3, s_4 and merging c-states s_5, s_6, see Table 1(3)

There could be two types of choices for getting $w^+(x)$ and $w^-(x)$. One based on how well x belongs to the corresponding state. A weight tends to be small if x marginally belongs to the state (e.g., locating near the boundary of S) but large if x firmly belongs to the state (e.g., locating deep inside S). The other bases on distributions $p(x|+)$ of red samples and $p(x|-)$ of green samples may be given by nonparamatric kernel estimation as follows:

Table 1 E-GPS states and E-GPS approach

Term	Description
	(1) ***Identification of system status by E-GPS states***
(a) E-GPS state	It is a convex set $R_s \subseteq \mathcal{D}_{g\phi e}$ with all its elements sharing the same status type, e.g., the state s_{11} in Fig. 1f, and the probability that the system visits this state (i.e., within R_s) is bigger than a threshold, i.e., the state is not rare. Empirically, the percentage of a given set of samples falling in R_s should be larger enough
(b) Prob. E-GPS state (*d-state* vs. *c-state*)	It is a state that is not rare but prob. (probabilistic) in a sense that each element in R_s is either Type '$+$' in a number n_s^+ or Type '$-$' in a number n_s^-, in two categories: d-state (Dedicated state): $max\{n_s^+, n_s^-\}$ is significantly bigger than $min\{n_s^+, n_s^-\}$. Empirically, samples falling in R_s are mostly dedicated to a same status type, e.g., the state s_{00} in Fig. 1f. c-state (Confusing state); otherwise, i.e., two status types compete samples in R_s, e.g., the states s_{10} and s_{01} in Fig. 1f
(c) c-state (cuttable vs noncuttable)	It is a c-state with at least one convex subset that is able to be cut off as a d-state, e.g., the state s_{01} in Fig. 1f; otherwise the c-state is said to be noncuttable under the current settings of $\mathcal{D}_{g\phi e}$, e.g., the state s_{10} in Fig. 1f
(d) Learning configuration of states	Overall, a set of at least one d-state and c-states (if any) is learned from a given set of samples, featured by not only these states but also their configuration that encodes the locations and mutual relations of these states, as illustrated in Fig. 1h
	(2) ***Refinements of E-GPS states***
(a) cutting	Cut a cuttable c-state by linear separation, e.g., SVM (Suykens and Vandewalle 1999; Suykens et al. 2002) or FDA by Eqs. (11) and (12) in Ref. Xu (2015a), via refining condition, e.g., the red line cuts s_{01} in Fig. 2f, which results in one convex subset as a d-state and one size-reduced c-state that may be still cuttable c-state
(b) merging	Merge adjacent d-states if their union is still convex, e.g., merging s_2, s_3, s_4 in Fig. 2f. In addition, merge adjacent c-states, e.g., s_5, s_6 in Fig. 2f
(c) growing	Grow each d-state s with $Type(s)$ ='green' by including those adjacent 'green' samples if the enlarged subset is still convex, and also grow each d-state s with $Type(s)$ ='red' by including those adjacent 'red' samples if the enlarged subset is still convex
(d) treating	Use additional conditions (e.g., one more variable is added to ϕ) such that more 'green' samples in the c-states become adjacent to and able to be re-allocated into some d-states in the above ways
	(3) ***Conditional phenotype analyses based on E-GPS states***
(a) analysis per d-state	*Prognosis analyses* test whether $max\{n_s^+, n_s^-\}$ differs from $min\{n_s^+, n_s^-\}$ significantly by χ^2 test or Fisher exact test to identify whether this state is good for prognosis, while the boundary of this state indicates the conditions under which the judgement is made. Moreover, prognosis of a unlabelled sample may be made by an one-class classifier obtained from these conditions *Survival analyses* plot K-M curves on samples with survival record and make the log rank test or the Cox proportional hazards test *Subtype analyses* stratify samples of this state into each subtype, test the enrichment of each subtype in this state, plot K-M curves on each stratification, and examine the correlation or the intersection of each subtype to good and bad prognosis, as shown in Fig. 1h
(b) analysis cross d-states	*Differentiation* test on whether there is a significant difference pair-wisely either between samples of different d-states or between samples associated with different values of a phenotype, in one of the following manners: ∗ A *t*-test when we ignore *e* and merely consider a univariate *g*; ∗ A multivariate test, e.g., Hotelling test Hotelling (1931), BBT test [see Table 6 in Ref. Xu (2015a)], and property-oriented test [see Algorithm 1 in Ref. Xu (2016)]; ∗ Model-based test proposed by Eqs. (29–31) in Ref. Xu (2015a);

Table 1 continued

Term	Description
	$*$ Logistic- or Cox-regression. On the lefthand of $\eta(\phi_t) = \mathbf{b}^T g_t + \mathbf{a}^T e_t + c + \varepsilon_t$, we test whether one or more of coefficients of **b** are zero and whether one or more of coefficients of **a** (e.g., by the score test or the Wald test) to examine whether the corresponding variables take roles significantly
	Staging that is related to subtypes but different, staging involves subtypes in a temporal order. The later stage is usually more serious than the earlier stage, which may be learned via the transfer probabilities $p(s_i\|s_j)$ cross the states in Fig. 1h
	Cross-state integration by comparing the configuration of states to enhance the differentiation study above. Moreover, cross-state combination can further provide better performance, as illustrated in Fig. 1h. Given the output measure $\zeta_{j,t}$ (e.g., p value, classification error, and predicted regression) for a particular sample t, we may get one weighted average $\zeta_t = \sum_j \zeta_{j,t} p(s_j\|t)$, as well as a combined classification rule $p(+\|t) = \sum_j p(+\|s_j)p(s_j\|t) > p(-\|t) = \sum_j p(-\|s_j)p(s_j\|t)$

$$
\begin{aligned}
p(x|+) &= \frac{1}{h^d n^+} \sum_{\text{each red sample } \xi} K\left(\frac{x-\xi}{h}\right),\\
p(x|-) &= \frac{1}{h^d n^-} \sum_{\text{each green sample } \xi} K\left(\frac{x-\xi}{h}\right),
\end{aligned}
\tag{9}
$$

where n^+, n^- are the total number of red and green samples, respectively, d is the dimension of x, and $h > 0$ is a small smoothing parameter. One simple example of $K(\frac{x-\xi}{h})$ is a Gaussian distribution with its mean ξ and the covariance $h^d I$.

As summarised in Table 1, the E-GPS approach is featured by identifying the system status via the E-GPS states that are learned from a given set of samples as addressed previously in this section and then further refined cutting, merging, growing as addressed in Table 1(2). Subsequently, we conduct various conditional phenotype analyses based on the E-GPS states, as summarised in Table 1(3).

Discussions

The E-GPS approach may find many uses in genomic biomarkers and cancer genetics, of which several applications are summarised in Table 2, including not only expression analyses and transcriptomic analysis of mRNA, lncRNA, and circRNA but also whole genome sequencing-based joint SNV analyses, mutation analyses, and methylation analyses, etc.

Additionally, it is also interesting to notice those degenerated situations with phenotype information unknown, e.g., all the red or green coloured points are turned into black dots. In such cases, all the states are degenerated into a same type, namely a unknown state or shortly called U-state. Each U-state actually represents a cluster of samples without any label information, and the task of identifying states is degenerated into clustering analysis, for which one possible method is learning a mixture of multiple local subspaces, e.g., see Algorithm 5 in Ref. Xu (2015b). In addition, we may consider

Table 2 Potential applications

Task	Study description
(a) Expression data differentiation	We find d-states as biomarkers by examining one g_a vs e_a (e.g., one g_a or g_b) by 2D scattering map. Also, one e_c can be jointly examined with one map for $e_c = 1$and one map for $e_c = 0$
(b) Mutation analysis	We examine one g_D vs e_c. First, get a 2×2 table for g_D. Then, the table is split into a 3D one with one slice for $e_c = 1$and the other for $e_c = 0$. Also, we may use one additional g_D as e_c to get a 3D table. Moreover, each slice may be further split by considering a new e_c. All the resulted slices are analysed in a way similar to Table 1(3)(a)
(c) SNP analysis	The situation is similar to the above except that a 2×2 table becomes a 2×3 table in consideration of g_D in a tri-nary values to denote AA, Aa, and aa. When using another SNP as e_c, its tri-valued g_D is replaced by a binary one that takes either 0 if the sample has no SNP on this site or 1 otherwise
(f) High-risk samples	Based on the above studies, we estimate the posteriori $p(+\mid x)$ per sample x and pick one with its value higher than a threshold as a high-risk sample, which is directly applicable to expression data. For sequencing data and particularly for finding SNPs, it difficult to get $p(+\mid x)$ because merely a few samples have variants on a particular site of g_c. Instead, a sample is regarded as risk simply when there is a variant on the site of g_c or an enough number of variants on the sites of multiple SNPs
(g) Expression-sequencing echoing	We obtain d-states and trees on expression data and sequencing data, and examine whether the results from two types of data in accordance with each other.
(h) Expression-sequencing combining (ESC) test	Assume the null H_0 holds on both the E-side and the S-side and using $E_{\neg H*}$ and $S_{\neg H*}$ to denote making alarm on its corresponding side, we get $p(E_{\neg H*}, S_{\neg H*} \mid s) = p(E_{\neg H*} \mid S_{\neg H*} \mid s) p_S$ with $p_S = p(S_{\neg H*} \mid s)$ being the p value obtained on the S-side and $p(E_{\neg H*} \mid S_{\neg H*}, s) \approx Card(B_E)/Card(B_S)$, being the probability of rejecting H_0 on the E-side conditioning on that H_0 is rejected on the S-side, where B_S consists of biomarkers on which H_0 is rejected significantly on the S-side, and $B_E \subseteq B_S$ consists of biomarkers on which H_0 is also regarded as significantly rejected on the S-side
(i) E-GPS based Integration	Integration may also be made by examining one g_a from expression of a gene versus g_c from multiple SNPs within the DNA sequence of the gene (e.g., either the number of or the average score of multiple SNPs)

**** General settings***

g: *each of its elements is a g-variable that could be*

g_a a real variable for expression of an RNA unit, e.g., either of mRNA, lncRNA, and circRNA;

g_b a real variable for a signature expression (i.e., a collective expression of a set of RNA-units);

g_c a discrete label for an SNP in DNA sequence (could be multiple SNPs per an RNA unit);

g_D a binary variable that indicates whether there is a mutation within a bio-unit sequence (e.g., gene, pathway, etc). There are usually multiple variables for different type mutations

$\boldsymbol{\phi}$: *each of its elements is a $\boldsymbol{\phi}$-variable that could be*

$\boldsymbol{\phi}_a$ a binary variable that indicates 'case vs control' or 'abnormal vs normal';

$\boldsymbol{\phi}_b$ a binary or discrete variable that indicates clinical features;

$\boldsymbol{\phi}_c$a discrete label that indicates one of subtypes or grades or stages;

$\boldsymbol{\phi}_D$ a real variable that indicates the occurrence of an event (e.g., survival time)

e: *each of its elements is an e-variable that could be*

e_a a g-variable that acts as a condition for our examination;

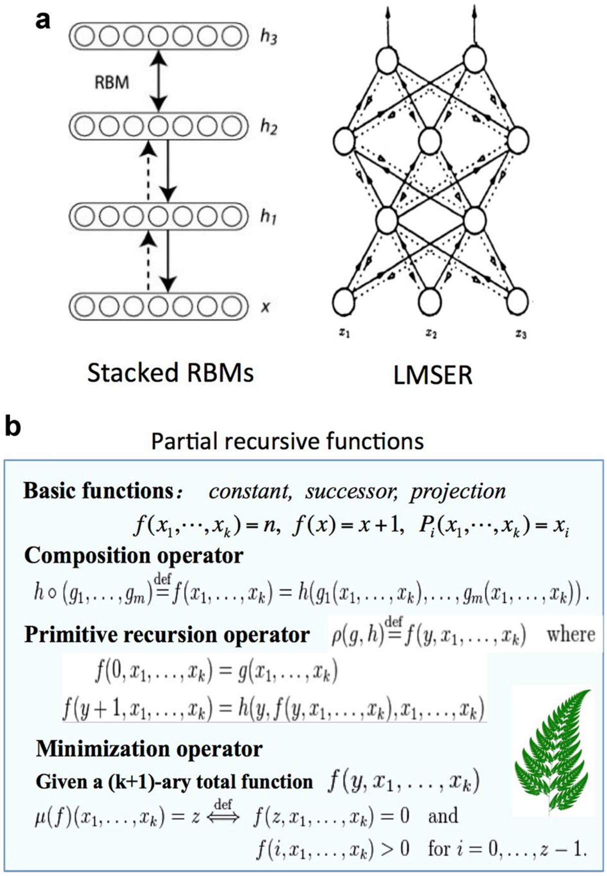
e_b a $\boldsymbol{\phi}$-variable that act as a condition for our examination;

e_c a binary variable that indicates whether a treatment is made, e.g., adjuvant chemotherapy;

e_D an environmental variable, in either discrete (e.g., sex M/F) or real (e.g., age) the binary factor analysis that describes 2^m states with the number of free parameters significantly reduced, e.g., see Algorithm 6 and Algorithm 7 in Ref. Xu (2015b).

Without considering phenotype information, the task nature will become really different from the E-GPS approach, in which phenotype analysis takes a core role in various tasks. We may use unsupervised learning as a preprocessing stage and the resulted clusters as one initial state configuration, on which the E-GPS study is further performed to take phenotype information in consideration.

In many biomarker searching tasks, the data may be mixed up by samples with phenotypes available and samples with phenotype unknown or partially missing. Considering unlabelled data may help to improve performances, which relates to semi-supervised



Basic functions: *constant, successor, projection*

$f(x_1,\cdots,x_k)=n,\ \ f(x)=x+1,\ \ P_i(x_1,\cdots,x_k)=x_i$

Composition operator

$h\circ(g_1,\ldots,g_m)\overset{\text{def}}{=}f(x_1,\ldots,x_k)=h(g_1(x_1,\ldots,x_k),\ldots,g_m(x_1,\ldots,x_k))\,.$

Primitive recursion operator $\rho(g,h)\overset{\text{def}}{=}f(y,x_1,\ldots,x_k)$ where

$f(0,x_1,\ldots,x_k)=g(x_1,\ldots,x_k)$

$f(y+1,x_1,\ldots,x_k)=h(y,f(y,x_1,\ldots,x_k),x_1,\ldots,x_k)$

Minimization operator

Given a (k+1)-ary total function $f(y,x_1,\ldots,x_k)$

$\mu(f)(x_1,\ldots,x_k)=z\overset{\text{def}}{\iff}f(z,x_1,\ldots,x_k)=0$ and

$f(i,x_1,\ldots,x_k)>0$ for $i=0,\ldots,z-1.$

Fig. 3 A possible extension: getting E-GPS biomarkers by deep learning. **a** Considering many genes and multiple conditional measures in a bio-system (e.g., a pathway) that consists of far more than a few genes, we may consider a multiple-layer network by deep learning, such as stacked RBMs (Hinton and Salakhutdinov 2006) and LMSER (Xu 1991, 1993) featured by unsupervised learning for a hierarchical abstraction of biomarkers. **b** According to Turing–Church thesis, the class of partial recursive functions is precisely the functions that can be computed by Turing machines, which provide an interesting perspective for understanding deep learning. The basic functions are involved within each layer, and the operators of composition and primitive recursion correspond a forward processing across different layers (namely what is usually called 'deep'), while the minimisation operator corresponds a recurrent process from upper layers back to lower layers. From this perspective, we speculate that the class of functions performed by deep neural networks is also the class of functions that can be computed by Turing machines, for which 'deep' and 'recurrent' are indispensable

learning, e.g., see Algorithm 9 in Ref. Xu (2015b) for semi-supervised clustering and Algorithm 11 in Ref. Xu (2015b) for semi-supervised binary factor analysis.

Another possible extension is getting the E-GPS biomarkers by deep learning multiple-layer networks, especially when we consider many genes and multiple conditional measures in a bio-system (e.g., pathway) that consists of far larger than a few genes. As illustrated in Fig. 3a, examples include stacked restricted Boltzmann machines (RBMs) (Hinton and Salakhutdinov 2006) and Least mean square error reconstruction (LMSER) (Xu 1991, 1993). Interestingly, the class of functions performed by deep neural networks is here speculated to be equivalently the class of functions that can be computed by Turing machines, from the perspective of partial recursive functions.

Conclusion

In the joint domain $\mathcal{D}_{g\phi e}$ of geno-measures, pheno-measures, and enviro-measures, those elements that locate adjacently in a convex subset are identified as forming a state as biomarkers. In place of a conventional biomarker that uses one or multiple g-measures as a biomarker unconditionally, this E-GPS approach provides a new biomarker analysis tool that considers not only geno-variables conditionally on certain focused domain but also the joint enviro-geno-pheno effect, as well as the E-GPS state based phenotype analyses such as differentiation, prognosis, subtype, staging, and pathogenic progression. Specifically, a two-stage method is proposed for learning these E-GPS states, and several possible applications are suggested. Moreover, it is further addressed that such an E-GPS approach facilitates integrative study of expression and sequencing.

Author details

[1] Department of Computer Science and Engineering, Centre for Brain-inspired Computing and Bio-Health Informatics, The School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, SEIEE Building 3, 800 Dongchuan Road, Minhang District, 200240 Shanghai, China. [2] Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong, China.

Acknowledgements

This work was supported by the Zhi-Yuan chair professorship start-up Grant from Shanghai Jiao Tong University.

Competing interests

The authors declare that they have no competing interests.

Received: 17 April 2016 Accepted: 20 July 2016

References

Bai H, Harmancı AS, Erson-Omay EZ, Li J, Coşkun S, Simon M, Krischek B, Özduman K, Omay SB, Sorensen EA (2016) Integrated genomic characterization of idh1-mutant glioma malignant progression. Nat Genet 48(1):59–66

Cristescu R, Lee J, Nebozhyn M, Kim K-M, Ting JC, Wong SS, Liu J, Yue YG, Wang J, Yu K (2015) Molecular analysis of gastric cancer identifies subtypes associated with distinct clinical outcomes. Nat Med 21(5):449–456

Dalerba P, Sahoo D, Paik S, Guo X, Yothers G, Song N, Wilcox-Fogel N, Forgó E, Rajendran PS, Miranda SP (2016) Cdx2 as a prognostic biomarker in stage II and stage III colon cancer. N Engl J Med 374(3):211–222

Hinton GE, Salakhutdinov RR (2006) Reducing the dimensionality of data with neural networks. Science 313(5786):504–507

Hotelling H (1931) The generalization of student's ratio. Ann Math Stat 2(3):360–378

Suykens JA, Vandewalle J (1999) Least squares support vector machine classifiers. Neural Process Lett 9(3):293–300

Suykens JA, Van Gestel T, De Brabanter J, De Moor B, Vandewalle J, Suykens J, Van Gestel T (2002) Least squares support vector machines. World Scientific Publishing, Singapore

Xu L (1991) Least mse reconstruction for self-organization:(i) multi-layer neural nets and (ii) further theoretical and experimental studies on one layer nets. In: Proceedings of the international joint conference on neural networks-1991-Singapore. pp 2363–2373

Xu L (1993) Least mean square error reconstruction principle for self-organizing neural-nets. Neural Netw 6(5):627–648

Xu L (2015a) Bi-linear matrix-variate analyses, integrative hypothesis tests, and case–control studies. Appl Inform 2(1):1–39

Xu L (2015b) Further advances on bayesian ying yang harmony learning. Appl Inform 2(5):1–45

Xu L (2016) A new multivariate test formulation: theory, implementation, and applications to genome-scale sequencing and expression. Appl Inform 3(1):1–23