Semantic distillation: a method for clustering objects by their contextual specificity

Thomas Sierocinski, Anthony Le Béchec, Nathalie Théret, Dimitri Petritis

Semantic distillation: a method for clustering objects by their contextual specificity

Thomas Sierocinski[1], Antony Le Béchec[2], Nathalie Théret[2], and Dimitri Petritis[1]

[1] Institut de recherche mathématique (UMR6625), Université de Rennes 1
[2] INSERM U620, Université de Rennes 1

Summary. Techniques for data-mining, latent semantic analysis, contextual search of databases, etc. have long ago been developed by computer scientists working on information retrieval (IR). Experimental scientists, from all disciplines, having to analyse large collections of raw experimental data (astronomical, physical, biological, etc.) have developed powerful methods for their statistical analysis and for clustering, categorising, and classifying objects. Finally, physicists have developed a theory of quantum measurement, unifying the logical, algebraic, and probabilistic aspects of queries into a single formalism.

The purpose of this paper is twofold: first to show that when formulated at an abstract level, problems from IR, from statistical data analysis, and from physical measurement theories are very similar and hence can profitably be cross-fertilised, and, secondly, to propose a novel method of fuzzy hierarchical clustering, termed *semantic distillation* — strongly inspired from the theory of quantum measurement —, we developed to analyse raw data coming from various types of experiments on DNA arrays. We illustrate the method by analysing DNA arrays experiments and clustering the genes of the array according to their specificity.

Keywords: Quantum information retrieval, semantic distillation, DNA microarray, quantum and fuzzy logic

1 Introduction

Sequencing the genome constituted a culminating point in the analytic approach of Biology. Now starts the era of the synthetic approach in Systems Biology where interactions among genes induce their differential expression that leads to the functional specificity of cells, the coherent organisation of cells into tissues, organs, and finally organisms.

However, we are yet far from a complete explanatory theory of living matter. It is therefore important to establish precise and quantitative phenomenology before being able to formulate a theory. The contribution of this paper is to provide the reader with a novel algorithmic method, termed *semantic distillation*, to analyse DNA arrays experiments (where genes are hybridised with various cell lines corresponding to various tissues or specific individuals) by determining the degree of specificity of

every gene to the particular context. The method provides experimental biologists with lists of candidate genes (ordered by their degree of specificity) for every biological context, clinicians with improved tools for diagnosis, pharmacologists with patient-tailored therapies, etc.

In the sequel we present the method split into several algorithmic tasks thought as subroutines of the general algorithm. It is worth noting that the method, although can profitably exploit, does not rely on any previous information stored in the existing databases; its rationale is to help analysing raw experimental data even in the absence of any previous knowledge.

The main idea of the method is summarised as follows. Experimental information hold on the objects of the system undergoes a sequence of processing steps; each step is performed on a different representation of the information. Those different representation spaces and the corresponding information processing act as successive filters revealing at the end the most pertinent and significant part of the information, hence the name "semantic distillation".

At the first stage, raw experimental data, containing all available information, are represented in an abstract Hilbert space, *the space of concepts* — reminiscent of the space of pure states in Quantum Mechanics —, endowing the set of objects with a metric space structure that is exploited to quantify the interactions among objects and encode them into a weighed graph on the vertex set of objects and with object interactions as edge weights.

Now objects (genes) are parts of an organised system (cell, tissue, organism). Therefore their mutual interactions are not just independent random variables; they are interconnected through precise, although certainly very complicated and mostly unknown relationships. We seek to reveal (hidden and unknown) interactions among genes. This is achieved by trading the weighed graph representation for a low-dimensional representation and using spectral properties of the weighed Laplacian on the graph to grasp the essential interactions.

The following step consists in a fuzzy divisive clustering of objects among two subsets by exploiting the previous low-dimensional representation. This procedure assigns a fuzzy membership to each object relative to characters of the two subsets. Fuzziness is as a matter of fact a distinctive property of experimental biological data reflecting our incomplete knowledge of fundamental biological processes.

Up to this step, our method is a sequence of known algorithms that have been previously used separately in the literature in various contexts. The novelty of our method relies on the following steps. The previous fuzzy clustering reduced the indeterminacy of the system. This information is fed back to the system to perform a projection to a proper Hilbert subspace. In that way, the information content of the dataset is modified by the information gained by the previous observations. After this feeding back, the three previous steps are repeated but now referring to a Hilbert spaces of lower dimension. Therefore our method is not a mere fuzzy clustering algorithm but a genuine non-classical interaction information retrieval procedure where previous observations alter the informational content of the system, reminiscent of the measurement procedure in Quantum Mechanics.

2 A Hilbert space formulation

2.1 Mathematical form of the dataset

Let $\mathbb{B}$ be a finite set of *documents* (or objects, or books) and $\mathbb{A}$ a finite set of *attributes* (or contexts, or keywords). The dataset is a $|\mathbb{B}| \times |\mathbb{A}|$ matrix $\mathsf{X} = (x_{ba})_{b\in\mathbb{B},a\in\mathbb{A}}$ of real or complex elements, where $|\cdot|$ represents cardinality. Equivalent ways of representing the dataset are

- a collection of $|\mathbb{B}|$ row vectors $\mathbf{x}_b = (x_{b1},\ldots,x_{b|\mathbb{A}|}), b \in \mathbb{B}$ of $\mathbb{R}^{|\mathbb{A}|}$ (or $\mathbb{C}^{|\mathbb{A}|}$),
- a collection of $|\mathbb{A}|$ column vectors $\mathbf{x}^a = (x_{1a},\ldots,x_{|\mathbb{B}|a}), a \in \mathbb{A}$ of $\mathbb{R}^{|\mathbb{B}|}$ (or $\mathbb{C}^{|\mathbb{B}|}$).

Example 1. In the experiments we analysed $\mathbb{B}$ is a set of 12000 human genes and $\mathbb{A}$ a set of 12 tissular contexts. The matrix elements x_{ba} are real numbers encoding luminescence intensities (or their logarithms) of DNA array ultimately representing the level of expression of gene b in context a.

Example 2. Let $\mathbb{B}$ be a set of books in a library and $\mathbb{A}$ a set of bibliographic keywords. The matrix elements x_{ba} can be $\{0,1\}$-valued: if the term a is present in the book b then $x_{ba} = 1$ else $x_{ba} = 0$. A variant of this example is when x_{ba} are integer valued: if the term a appears k times in document b then $x_{ba} = k$.

Example 3. Let $\mathbb{B}$ be a set of students and $\mathbb{A}$ a set of papers they gave. The matrix elements x_{ba} are real valued; x_{ba} is the mark the student b got in paper a.

The previous examples demonstrate the versatility of the method by keeping the formalism at an abstract level to apply indistinctively into various very different situations without any change. Note also that the assignment as set of documents or attributes is a matter of point of view; for instance, example 3 as it stands is convenient in evaluating students. Interchanging the role of sets $\mathbb{A}$ and $\mathbb{B}$ renders it adapted to the evaluation of teaching. As a rule of thumb, in biological applications, $|\mathbb{A}| \ll |\mathbb{B}|$.

2.2 The space of concepts

For $\mathbb{A}$ and $\mathbb{B}$ as in the previous subsection, we define the *space of concepts*, $\mathscr{H}_{\mathbb{A}}$, as the real or complex free vector space over $\mathbb{A}$, i.e. elements of $\mathbb{A}$ serve as indices of an orthonormal basis of $\mathscr{H}_{\mathbb{A}}$. Therefore, the complete dataset X can be represented as the collection of $|\mathbb{B}|$ vectors $|\,\Xi_b\,\rangle = \sum_{a\in\mathbb{A}} x_{ba}|\,a\,\rangle \in \mathscr{H}_{\mathbb{A}}$, with $b \in \mathbb{B}$ and where $|\,a\,\rangle$ represents the element of the orthonormal basis of the free vector space corresponding to the attribute a. We use here Dirac's notation to represent vectors, linear forms and projectors on this space (see any book on quantum mechanics or [26] for a freely accessible document and [29] for the use of this notation in information retrieval). The vector $|\,\Xi_b\,\rangle$ contains all available experimental information on document b in various cellular contexts indexed by the attributes a; it can be thought as a convenient bookkeeping device of the data $(x_{ba})_{a\in\mathbb{A}}$, in the same way a generating function contains all the information on a sequence as formal power series.

The vector space is equipped with a scalar product defined for every two vectors $|\psi\rangle = \sum_{a\in\mathbb{A}} \psi_a |a\rangle$ and $|\psi'\rangle = \sum_{a\in\mathbb{A}} \psi'_a |a\rangle$ by $\langle\psi|\psi'\rangle = \sum_{a\in\mathbb{A}} \overline{\psi}_a \psi'_a$, where $\overline{\psi}_a$ denotes the complex conjugate of ψ_a (it coincides with ψ_a if it is real). Equipped with this scalar product, the vector space $\mathscr{H}_{\mathbb{A}}$ becomes a real or complex $|\mathbb{A}|$-dimensional Hilbert space. The scalar product induces a Hilbert norm on the space, denoted by $\|\cdot\|$. In the sequel we introduce also *rays* on the Hilbert space i.e. normalised vectors. Since the dataset X does not in principle verify any particular numerical constraints, rays are constructed by dividing vectors by their norms. We use the symbol $|\xi_b\rangle = |\Xi_b\rangle / \||\Xi_b\rangle\|$ to denote the ray associated with vector $|\Xi_b\rangle$.

The Hilbert space structure on $\mathscr{H}_{\mathbb{A}}$ allows a natural geometrisation of the space of documents by equipping it with a pseudo-distance[3] $d : \mathbb{B}\times\mathbb{B} \to \mathbb{R}_+$ defined by $d(b,b') = \||\xi_b\rangle - |\xi_{b'}\rangle\|$. What is important here is not the precise form of the pseudo-metric structure of $(\mathbb{B}, d)$; several other pseudo-distances can be introduced, not necessarily compatible with the scalar product. In this paper we stick however to the previous pseudo-distance, postponing into a later publication explanations about the significance of other pseudo-distances.

As is the case in Quantum Mechanics, the Hilbert space description incorporates into a unified algebraic framework all logical and probabilistic information hold by the dataset. An enquiry of the type "does the system possess feature F" is encoded into a projector P_F acting on the Hilbert space. The subspace associated with the projector P_F is interpreted as the set of documents retrieved by asking the question about the feature F. Now all experimental information hold by the dataset is encoded into the *state* of the system represented by a *density matrix* ρ (i.e. a self-adjoint, positive, trace class operator acting on $\mathscr{H}_{\mathbb{A}}$ having unit trace). Retrieved documents possess the feature F with probability $\mathtt{tr}(\rho P_F)$. Thus the algebraic description incorporates logical information on the documents retrieved as relevant to a given feature and assign them a probability determined by the state defined by the experiment. For example, the probability that a gene b is relevant to an attribute a is given by the above formula with $P = |a\rangle\langle a|$ and $\rho = |\xi_b\rangle\langle\xi_b|$, yielding $\mathtt{tr}(\rho P) = |\langle\xi_b|a\rangle|^2$.

3 A weighed graph with augmented vertex set

The careful reader has certainly already noted that in the above description vectors $|\xi_b\rangle$, encoding the information about document b, and basis vectors $|a\rangle$, associated with attribute a, all belong to the same Hilbert space $\mathscr{H}_{\mathbb{A}}$. Therefore, although initially the sets $\mathbb{A}$ and $\mathbb{B}$ are disjoint since they have distinct elements, when passing to the Hilbert space representation, vectors $|\xi_b\rangle$ and $|a\rangle$ have very similar roles in representing indistinguishably objects or attributes as vectors of $\mathscr{H}_{\mathbb{A}}$. In the sequel, we introduce the set $\mathbb{V}$ (or more precisely $\mathbb{V}_{\mathbb{A}}$ to remove any ambiguity) as the set $\mathbb{V}_{\mathbb{A}} = \mathbb{A}\cup\mathbb{B}$. Thus, for any $v \in \mathbb{V}_{\mathbb{A}}$,

[3] It is termed pseudo-distance since it verifies symmetry and triangle inequality but $d(b,b')$ can vanish even for different b and b'. As a matter of fact, d is a distance on the projective Hilbert space.

$$|\,\Xi_v\,\rangle = \begin{cases} |\,a\,\rangle & \text{if } v = a \in \mathbb{A}, \\ \sum_{a\in\mathbb{A}} x_{ba}|\,a\,\rangle & \text{if } v = b \in \mathbb{B}. \end{cases}$$

The new vectors $|\,\Xi_a\,\rangle = |\,a\,\rangle$ are included as *specificity witnesses* in the dataset. Note that since these new vectors are also elements of the same Hilbert space, the pseudo-distance d naturally extends to $\mathbb{V}_{\mathbb{A}}$.

Suppose now that a *similarity function* $\sigma : \mathbb{V}_{\mathbb{A}} \times \mathbb{V}_{\mathbb{A}} \to [0,1]$ is defined. For the sake of definiteness, the reader can think of σ as being given, for example[4], by $\sigma(v,v') = \sqrt{1-\frac{1}{2}d(v,v')^2}$; results we quote in section 5 are obtained with a slight modification of this similarity function. However, again, the precise form of the similarity function is irrelevant in the abstract setting serving as foundation of the method. Several other similarity functions have been used like, for example, $\sigma(v,v') = \exp(-\|\Xi_v - \Xi_{v'}\|^2/\tau)$ with τ a positive constant or some others, in particular, functions taking value 0 even for some vertices corresponding to non orthogonal rays but the explanation of their significance is postponed to a subsequent publication.

A weighed graph is now constructed with vertex set $\mathbb{V}_{\mathbb{A}}$. Weights are assigned to the edges of the complete graph over $\mathbb{V}_{\mathbb{A}}$; the weights being expressible in terms of the similarity function σ. Again, the precise expression is irrelevant for the exposition of the method. For the sake of concreteness, the reader can suppose that the weights $W_{vv'}$ are given by $W_{vv'} = \sigma(v,v')$. The pair $(\mathbb{V}_{\mathbb{A}}, W)$ with W being the symmetric matrix $W = (W_{vv'})_{v,v\in\mathbb{V}_{\mathbb{A}}}$, denotes the weighed graph.

At this level of the description we follow now standard techniques of reduction of the data dimensionality by optimal representation of the graph in low dimensional Euclidean spaces spanned by eigenvectors of the Laplacian. Such methods have been used by several authors [4, 24]. Here we give only the basic definitions and main results of this method. The interested reader may consult standard textbooks like [8, 10, 15] for general exposition of the method.

Definition 1. *A map $\mathbf{r} : \mathbb{V}_{\mathbb{A}} \to \mathbb{R}^{\nu}$ is called a ν-dimensional representation of the graph. The representation is always supposed non-trivial (i.e. $\mathbf{r} \not\equiv 0$) and balanced (i.e. $\sum_{v\in\mathbb{V}_{\mathbb{A}}} \mathbf{r}(v) = 0$).*

From the weights matrix W we construct the *weighed Laplacian matrix* $\Lambda = D - W$ where the matrix elements $D_{vv'}$ are 0 if $v \neq v'$ and equal to $\sum_{v''\in\mathbb{V}_{\mathbb{A}}} W_{vv''}$ if $v = v'$. More precisely, we denote by $\Lambda(\mathbb{V}_{\mathbb{A}})$ this weighed Laplacian to indicate that it is defined on the vertex set $\mathbb{V}_{\mathbb{A}}$. This precision will be necessary in the next section specifying the semantic distillation algorithm where the vertex set will be recursively modified at each step. The *weighed energy of the representation* is given by

$$\mathscr{E}_W(\mathbf{r}) = \sum_{v,v'\in\mathbb{V}_{\mathbb{A}}} W_{vv'}\|\mathbf{r}(v) - \mathbf{r}(v')\|^2,$$

where in this formula $\|\cdot\|$ denotes the Euclidean norm of $\mathbb{R}^{\nu}$.

[4] This function is well adapted to datasets $\mathsf{X} = (x_{ba})$, with $x_{ba} \in \mathbb{R}^+$; for more general datasets, the factor $1/2$ must be changed to $1/4$.

Theorem 1. *Let $N = |\mathbb{V}_\mathbb{A}|$ and $\{\lambda_1, \ldots, \lambda_N\}$ be the spectrum of Λ, ordered as $\lambda_1 \leq \lambda_2 \ldots \lambda_N$. Suppose that $\lambda_2 > 0$. Then $\inf_\mathbf{r} \mathscr{E}_W(\mathbf{r}) = \sum_{i=2}^{\nu+1} \lambda_i$, where the infimum is over all ν-dimensional non-trivial balanced representations of the graph.*

Remark 1. If $\mathbf{u}^1, \ldots \mathbf{u}^N$ are the eigenvectors of Λ corresponding to the eigenvalues $\lambda_1, \ldots, \lambda_N$ ordered as above, then $\mathbf{u}^2$ is the best one-dimensional, $[\mathbf{u}^2, \mathbf{u}^3]$ the best two-dimensional, etc, $[\mathbf{u}^2, \ldots, \mathbf{u}^{\nu+1}]$ the best ν-dimensional representation of the graph $(\mathbb{V}_\mathbb{A}, W)$.

4 Fuzzy semantic clustering and distillation

The algorithm of semantic distillation is a recursive divisive fuzzy clustering followed by a projection on a Hilbert subspace and a thinning of the graph. It starts with the Hilbert space $\mathscr{H}_\mathbb{A}$ and the graph with vertex set $\mathbb{V}_\mathbb{A}$ and constructs a sequence of Hilbert subspaces and subgraphs indexed by the words κ of finite length on a two-letter alphabet. This set is isomorphic to a subset of the rooted binary tree. If κ is the root, then define $\mathbb{M}_\kappa = \mathbb{A}$. Otherwise, $\mathbb{M}_\kappa$ will be a proper subset of $\mathbb{A}$, i.e. $\emptyset \subset \mathbb{M}_\kappa \subset \mathbb{A}$,indexed by κ. When $|\mathbb{M}_\kappa| = 1$ then the corresponding κ is a leaf of the binary tree. The algorithm stops when all indices correspond to leaves.

More precisely, let $\mathbb{K} = \{1, 2\}$, $\mathbb{K}^0 = \{\kappa : \kappa = ()\}$, and for integers $n \geq 1$ let $\mathbb{K}^n = \{\kappa : \kappa = \kappa_1 \cdots \kappa_n; \kappa_i \in \mathbb{K}\}$. Finally let $\mathbb{K}^* = \cup_{n \geq 0} \mathbb{K}^n$ denote the set of words on two letters of indefinite length, including the empty sequence, denoted by $()$, of zero length that coincides with the root of the tree. If $\kappa = \kappa_1 \cdots \kappa_n$ is a word of n letters and $k \in \mathbb{K}$, we denote the concatenation κk as the word of $n+1$ letters $\kappa_1 \cdots \kappa_n k$.

We start from the empty set $\mathsf{Leaves} = \{\}$, the empty sequence $\kappa = ()$ and the current attributes set $\mathbb{M}_\kappa = \mathbb{M}_{()} = \mathbb{A}$ and current tree $\mathsf{Tree} = \{\kappa\}$. We denote $\mathbb{V}_\kappa = \mathbb{B} \cup \mathbb{M}_\kappa$. We need further a *fuzzy membership* function $m : \mathbb{V}_\kappa \times \mathbb{K} \to [0, 1]$. The fuzzy clustering algorithm is succinctly described as Algorithm 1 below.

Data: κ, $\mathbb{M}_\kappa$, $\mathbf{r}$, objective function F
Result: Two sets $\mathbb{M}_{\kappa 1}$ and $\mathbb{M}_{\kappa 2}$ and the fuzzy membership $m(v,k)$ for $v \in \mathbb{V}_\kappa$ in the clusters $\mathbb{M}_{\kappa 1}$ and $\mathbb{M}_{\kappa 2}$
if $|\mathbb{M}_\kappa| > 1$ **then**
  assign $(v_1, v_2) \leftarrow \arg\max\{\|\mathbf{r}(v) - \mathbf{r}(v')\|, v, v' \in \mathbb{V}_\kappa\}$;
  assign $\mathbf{r}(v_1)$ and $\mathbf{r}(v_2)$ as centroids for the two candidate finer clusters $\mathbb{M}_{\kappa 1}$ and $\mathbb{M}_{\kappa 2}$;
  use standard 2-means fuzzy clustering algorithm to minimise objective function F under the constraint $\sum_{k=1}^{2} m(v,k) = 1$, for all $v \in \mathbb{M}_\kappa$;
  assign $\mathbb{M}_{\kappa 1} \leftarrow \{v \in \mathbb{M}_\kappa : m(v,1) > m(v,2)\}$;
  assign $\mathbb{M}_{\kappa 2} \leftarrow \mathbb{M}_\kappa \setminus \mathbb{M}_{\kappa 1}$;
end

Algorithm 1: FuzzyClustering

Note that in the previous construction $\mathbb{M}_{\kappa k} \subset \mathbb{M}_\kappa$ for every κ and every $k \in \mathbb{K}$. Therefore, the algorithm explores the branches of a tree from the root to the leaves. Denote by π_κ the orthogonal projection from $\mathscr{H}_\mathbb{A}$ to $\mathscr{H}_{\mathbb{M}_\kappa}$. The distillation step is described by the following Algorithm 2.

Data: FuzzyClustering
Result: Leaves and sequence of singleton sets $\mathbb{M}_\kappa$ for $\kappa \in$ Leaves
Initialisation{
$\kappa \leftarrow ()$;
$\mathbb{M}_\kappa \leftarrow \mathbb{A}$;
Leaf$(\kappa) \leftarrow \mathbb{M}\kappa$;
Leaves $\leftarrow \{\}$;
Tree $\leftarrow \{\kappa\}$;
Bookkeeping $\leftarrow \{\kappa\}$;
}
while *Bookkeeping* $\neq \emptyset$ **do**
  for $\kappa \in$ *Bookkeeping* **do**
    if $|\mathbb{M}_\kappa| = 1$ **then**
      Leaves $\leftarrow$ Leaves $\cup \{\kappa\}$;
      Bookkeeping $\leftarrow$ Bookkeeping $\setminus \{\kappa\}$;
    else
      Use π_κ to project from $\mathscr{H}_\mathbb{A}$ to $\mathscr{H}_{\mathbb{M}_\kappa}$;
      Thin the graph: $\mathbb{V}_\kappa \leftarrow \mathbb{B} \cup \mathbb{M}_\kappa$;
      Compute weighed Laplacian $\Lambda(\mathbb{V}_\kappa)$;
      Diagonalise $\Lambda(\mathbb{V}_\kappa)$;
      Compute ν-dimensional representation $\mathbf{r}$;
      Call FuzzyClustering;
      for $k \in \mathbb{K}$ **do**
        $\kappa' \leftarrow \kappa k$;
        Leaf$(\kappa') \leftarrow \mathbb{M}_{\kappa'}$ /* $\mathbb{M}_\kappa$ as determined by FuzzyClustering */;
        Tree $\leftarrow$ Tree $\cup \{\kappa'\}$;
        Bookkeeping $\leftarrow$ Bookkeeping $\cup \{\kappa'\}$;
      end
    end
  end
end

Algorithm 2: Distillation

5 Illustration of the method, robustness and complexity issues

We tested the method on a dataset for an experiment on DNA array published in [35], with the set $\mathbb{A}$ of attributes corresponding to 12 cell lines (bone marrow, liver, heart, spleen, lung, kidney, skeletal muscle, spinal cord, thymus, brain, prostate, pancreas) and the set $\mathbb{B}$ of documents corresponding to 12000 human genes. To illustrate the method we present here only an example of the type of results we obtain

by our method for the simplest case of one-dimensional representation of the graph. The complete lists of specificity degrees for the various genes (including their UniGene identifiers) for various dimensions are provided as supplemental material (at the home page of the first author).

Note that for one-dimensional representation, ordering by the magnitude of the eigenvector components is equivalent to a relabelling of genes. The figure 1 represents, within the previous mentioned relabelling, the levels of expressions for clustered genes. The same procedure has been applied for higher dimensional represen-

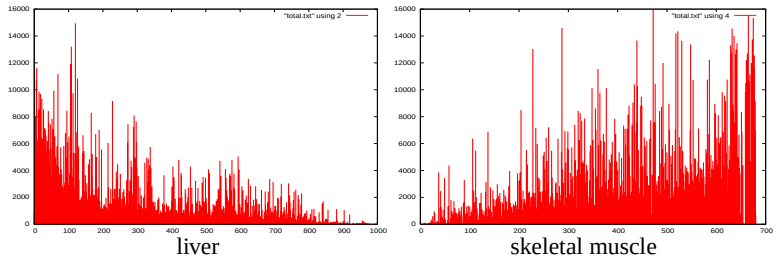

Fig. 1. For every singleton cluster, i.e. tissular context $\kappa \in \mathsf{Leaves}$ (we present solely the cases $\mathbb{M}_\kappa = \{\text{liver}\}$ and $\mathbb{M}_\kappa = \{\text{ skeletal muscle}\}$ in this example), the horizontal axis contains the set $\mathbb{B}$ of genes *relabelled* according to their decreasing (resp. increasing) fuzzy membership to $\mathbb{M}_\kappa$. Vertical axis represents the experimentally measured level of expression for those genes.

tation of the graph (i.e. $\nu > 1$). These results are not presented here; they marginally improved some specifications and helped us removing apparent degeneracy in some cases. Finally, in the table 1, we give an example of the annotation provided by the database UniGene for the genes classified as specific of skeletal muscle cell line by our method.

We observe that the majority of genes classified as most specific by our method are in fact annotated as specific in the database. To underline the power of our method, note that the UniGene annotation for the ATPase gene is “cardiac muscle”. Our method determines it as most specific of “skeletal muscle”. We checked the experimental data we worked on and realised that this gene is, as a matter of fact, 5 times more expressed in the skeletal muscle context than in the cardiac muscle. Therefore, our method correctly determines this gene as skeletal-muscle-specific.

In summarising, our method is an automatic and algorithmic method of analysis of raw experimental data; it can be used to any experiment of similar type *independently of any previous knowledge included in genomic databases* to provide biologists with a powerful tool of analysis. In particular, since most of the genes are not yet annotated in the existing databases, the method provides biologists with candidate genes for every particular context for further investigation. Moreover, the genetic character of documents and attributes is purely irrelevant; the same method

Table 1. Annotation of the genes closest (within the relabelling induced by $\mathbf{u}^2$) to the specificity witness "skeletal muscle". Genes are separated by the – symbol.

ATPase, Ca++ transporting, cardiac muscle, fast twitch 1, calcium signaling pathway – Troponin I type 2; skeletal, fast – Myosin, light chain 1, alkali; skeletal, fast – Ryanodine receptor 1; skeletal; calcium signaling pathway – Fructose-1,6-bisphosphatase 2, glycolysis / gluconeogenesis – Actinin, alpha 3; focal adhesion – Adenosine monophosphate deaminase 1 (isoform M) purine metabolism – Troponin C type 2; fast; calcium signaling pathway – Carbonic anhydrase III, muscle specific; nitrogen metabolism – Nebulin – Troponin I type 1; skeletal, slow – Myosin, heavy chain 3, skeletal muscle – Myogenic factor 6, herculin – Myosin binding protein C, fast type – Calcium channel, voltage-dependent, beta 1 subunit – Metallothionein 1X – Bridging integrator 1 – Bridging integrator 1 – Calpain 3, (p94) – Tropomyosin 3 – Phosphorylase, glycogen; muscle (McArdle syndrome, glycogen storage disease type V); starch and sucrose metabolism – Myozenin 3 – Myosin binding protein C, slow type – Troponin T type 3; skeletal, fast – Superoxide dismutase 2; mitochondrial – Nicotinamide N-methyltransferase – Sarcolipin – Interleukin 32 – Sodium channel, voltage-gated, type IV, alpha subunit – Guanidinoacetate N-methyltransferase; urea cycle and metabolism of amino groups.

can be used to any other dataset of similar structure, let them concern linguistic, genetic, or image data.

Concerning the algorithmic complexity of the method, the dominant contribution comes from the diagonalisation of a $|\mathbb{B}| \times |\mathbb{B}|$ dense real symmetric matrix, requiring at worst $\mathscr{O}(|\mathbb{B}|^3)$ time steps and $\mathscr{O}(|\mathbb{B}|^2)$ space. The time complexity can be slightly reduced, if only low-dimensional (dimension ν) representations are sought, to $\mathscr{O}(\nu \times |\mathbb{B}|^2)$ time steps. Moreover, we tested the method against additive or multiplicative random perturbations of the experimental data; it proved astonishingly robust.

6 Connections to previous work

The algorithm of semantic distillation maps the dataset into a graph and uses spectral methods and fuzzy clustering to analyse the graph properties. As such, this algorithm is inspired by various pre-existing algorithms and borrows several elements from them.

The oldest implicit use of a vector space structure to represent dataset and application of spectral methods to analyse them is certainly "principal components analysis" introduced in [25]. The method seeks finding directions of maximal variability in the space corresponding to linear combinations of the underlying vectors. The major drawbacks of principal components analysis are the assumptions that dataset matrix is composed of row vectors that are independent and identically distributed realisations of the same random vector (hence the covariance matrix whose principal components are sought can be approximated by the empirical covariance of the process) and that there exists a linear transformation maximising the variability.

Vector space representations and singular value decomposition, as reviewed in [5], have been used to retrieve information from digital libraries. Implementations of these ideas range from the famous PageRank algoritm used by Google (see [18]

and [17] for expository reviews) to whole genome analysis based on latent semantic indexing [23, 16].

From the information contained in the dataset X, a weighed graph of interactions among documents is constructed. To palliate the weaknesses of principal component analysis, reproducing kernel methods can be used. The oldest account of these methods seems to be [21] and their formulation in the context of Hilbert spaces can be found in [1]. In [31], analysis of features of a microarray experiment is proposed based on kernel estimates on a graph. Note however that in that paper, the graph incorporates extrinsic information coming from participation of genes in specific pathways as documented in the KEGG database. On the contrary, in the method we are proposing here, the graph can be constructed in an intrinsic way, even in the absence of any additional information from existing databases. In [4, 9, 24], kernel methods and Laplace eigenspace decomposition are used to generalise principal components analysis to include non-linear interactions among genes. Particular types of kernels, defined in terms of commuting times for a random walk on the graph are used in [13, 20, 30]. All these methods, although not always explicitly stated in these articles, are as a matter of fact very closely related since the kernels, the weighed graph Laplacian and the simple random walk on the graph can be described in a unified formalism [7, 8, 10, 15, 22]. It is worth noting that analysis of Laplacian of the graph is used in many different contexts, ranging from biological applications (proteins conformation [32], gene arrays [23]) to web search [3] or image analysis [28].

Fuzzy clustering has been introduced in [6]; lately it was shown [33] equivalent to probabilistic clustering if the objective function is expressed in terms of the Rényi entropy.

The idea of describing the data in terms of abstract Hilbert spaces has been used (in the context of database search) in [2, 12, 14, 29, 34].

The semantic distillation algorithm is based on a quantum-inspired subspace projection, strongly reminiscent of the quantum procedure of measurement. Although fully implemented on classical computers, it shares with general quantum algorithms features of non-distributive quantum logic [26, 27]. The semantic approach of Quantum Mechanics can be found in [27, 11]. It is worth underlying that the full fledged fuzzy logic induced by quantum semantics is not equivalent to the standard fuzzy logic introduced in [36]; it represents a genuine extension of it [11].

7 Perspectives

Various data sets (not only biological) are presently semantically distilled and the method compared with more traditional approaches. Preliminary results obtained so far seem to confirm the power of the method.

Several directions are in progress:

- Although the method is quantum-inspired, the fuzzy logic induced is still standard fuzzy logic. We are currently working on the extension to generalised fuzzy logic induced by full-fledged quantum semantics.

- The graph analysis we performed provided us with degrees of specificities of every gene in a particular context. These data can be reincorporated to the graph as internal degrees of freedom of a multi-layered graph that can be further analysed.
- The connections of the algorithm of semantic distillation with the algorithm of purification of quantum states [19] introduced in the context of quantum computing are currently explored.

References

1. N. Aronszajn. Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, 68:337–404, 1950.
2. Ricardo Baeza-Yates. Information retrieval in the web: beyond current search engines. *Internat. J. Approx. Reason.*, 34(2-3):97–104, 2003. Soft computing applications to intelligent information retrieval on the Internet (Mérida/Granada, 2002).
3. Pierre Baldi, Paolo Frasconi, and Padhraic Smyth. *Modeling the Internet and the Web: Probabilistic Methods and Algorithms*. Wiley InterScience, New York, 2003.
4. Mikhail Belkin and Partha Niyogi. Towards a theoretical foundation for Laplacian-based manifold methods. In *Learning theory*, volume 3559 of *Lecture Notes in Comput. Sci.*, pages 486–500. Springer, Berlin, 2005.
5. Michael W. Berry, Zlatko Drmač, and Elizabeth R. Jessup. Matrices, vector spaces, and information retrieval. *SIAM Rev.*, 41(2):335–362 (electronic), 1999.
6. James C. Bezdek. *Pattern recognition with fuzzy objective function algorithms*. Plenum Press, New York, 1981. With a foreword by L. A. Zadeh, Advanced Applications in Pattern Recognition.
7. Massimo Campanino and Dimitri Petritis. On the physical relevance of random walks: an example of random walks on a randomly oriented lattice. In *Random walks and geometry*, pages 393–411. Walter de Gruyter GmbH & Co. KG, Berlin, 2004.
8. Fan R. K. Chung. *Spectral graph theory*, volume 92 of *CBMS Regional Conference Series in Mathematics*. Published for the Conference Board of the Mathematical Sciences, Washington, DC, 1997.
9. Ronald R. Coifman and Stéphane Lafon. Diffusion maps. *Appl. Comput. Harmon. Anal.*, 21(1):5–30, 2006.
10. Dragoš M. Cvetković, Michael Doob, and Horst Sachs. *Spectra of graphs*. Johann Ambrosius Barth, Heidelberg, third edition, 1995. Theory and applications.
11. Maria Luisa Dalla Chiara, Roberto Giuntini, and Roberto Leporini. Compositional and holistic quantum computational semantics. *Natural Computing*, 6(5):113–132, 2007.
12. Sándor Dominich. *Mathematical foundations of information retrieval*, volume 12 of *Mathematical Modelling: Theory and Applications*. Kluwer Academic Publishers, Dordrecht, 2001.
13. Francois Fouss and Jean-Michel Renders. Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 19(3):355–369, 2007. Member-Alain Pirotte and Member-Marco Saerens.
14. Peter Gärdenfors. Induction, conceptual spaces and AI. *Philos. Sci.*, 57(1):78–95, 1990.
15. Chris Godsil and Gordon Royle. *Algebraic graph theory*, volume 207 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 2001.
16. Ramin Homayouni, Kevin Heinrich, Lai Wei, and Michael W. Berry. Gene clustering by latent semantic indexing of medline abstracts. *Bioinformatics*, 21(1):104–115, 2005.

17. Amy N. Langville and Carl D. Meyer. A survey of eigenvector methods for Web information retrieval. *SIAM Rev.*, 47(1):135–161 (electronic), 2005.
18. Amy N. Langville and Carl D. Meyer. A reordering for the PageRank problem. *SIAM J. Sci. Comput.*, 27(6):2112–2120 (electronic), 2006.
19. Hans Maassen and Burkhard Kümmerer. Purification of quantum trajectories. In *Dynamics & stochastics*, volume 48 of *IMS Lecture Notes Monogr. Ser.*, pages 252–261. Inst. Math. Statist., Beachwood, OH, 2006.
20. M. Meilă and J. Shi. A random walks view of spectral segmentation, In AI and Statistics, (2001).
21. J. Mercer. Functions of positive and negative type and their connection with the theory of integral equations. *Phil. Trans. Roy. Soc. London Ser. A*, 209:415–446, 1909.
22. Bojan Mohar. Some applications of Laplace eigenvalues of graphs. In *Graph symmetry (Montreal, PQ, 1996)*, volume 497 of *NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci.*, pages 225–275. Kluwer Acad. Publ., Dordrecht, 1997.
23. See-Kiong Ng, Zexuan Zhu, and Yew-Soon Ong. Whole-genome functional classification of genes by latent semantic analysis on microarray data. In *2nd Asia-Pacific Bioinformatics Conference, Dunedin, New Zealand*, volume 29 of *Conferences in research and practice in information technology*. Australian Computer Society, Inc., 2004.
24. Jens Nilsson. *Nonlinear dimensionality reduction of gene expression data*. PhD thesis, Lund University, Faculty of Engineering, 2005.
25. K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2:559–572, 1901.
26. Dimitri Petritis. Mathematical foundations of quantum mechanics and its applications, Preliminary version (2006) available at http://perso.univ-rennes1.fr/dimitri.petritis/ps/qiccc.pdf.
27. Miklós Rédei. *Quantum logic in algebraic approach*, volume 91 of *Fundamental Theories of Physics*. Kluwer Academic Publishers Group, Dordrecht, 1998.
28. Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):888–905, 2000.
29. C. J. van Rijsbergen. *The geometry of information retrieval*. Cambridge University Press, Cambridge, 2004.
30. Sebastien Vast, Pierre Dupont, and Yves Deville. Automatic extraction of relevant nodes in biochemical networks, citeseer.ist.psu.edu/742958.html.
31. J. Vert and M. Kanehisa. Graph-driven features extraction from microarray data using diffusion kernels and kernel cca, NIPS (2002). citeseer.ist.psu.edu/vert02graphdriven.html.
32. Saraswathi Vishveshwara, K. V. Brinda, and N. Kannan. Protein structure: insights from graph theory. *Journal of Theoretical and Computational Chemistry*, 1:187–212, 2002.
33. S. T. Wang, K. F. Chung, H. B. Shen, and R. Q. Zhu. Note on the relationship between probabilistic and fuzzy clustering. *Soft Computing*, 8:523–526, 2004.
34. Dominic Widdows. *Geometry and meaning*, volume 172 of *CSLI Lecture Notes*. CSLI Publications, Stanford, CA, 2004. With a foreword by Pentti Kanerva.
35. Itai Yanai, Hila Benjamin, Michael Shmoish, Vered Chalifa-Caspi, Maxim Shklar, Ron Ophir, Arren Bar-Even, Shirley Horn-Saban, Marilyn Safran, Eytan Domany, Doron Lancet, and Orit Shmueli. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics*, 21:650–659, 2005.
36. L. A. Zadeh. Fuzzy sets. *Information and Control*, 8:338–353, 1965.