Design and Applications of Intelligent Systems in Identifying Future Occurrence of Tuberculosis Infection in Population at Risk

Adel Ardalan[1], Ebru Selin Selen[2], Hesam Dashti[3], Adel Talaat[4], Amir Assadi[3]

[1] Department of Electrical and Computer Engineering
University of Wisconsin, Madison WI 53706 USA
[2]Department of Applied Statistics and Applied Mathematics,
Izmir University of Economics, 35330 Balcova, TR
[3] Department of Mathematics, University of Wisconsin,
Madison, WI 53706 Madison, WI 53706
[4] Department of Animal Health and Biomedical Sciences
University of Wisconsin, Madison WI 53706 USA

Abstract. Tuberculosis is a treatable but severe disease caused by Mycobacterium tuberculosis (Mtb). Recent statistics by international health organizations estimate the Mtb exposure to have reached over two billion individuals. Delay in disease diagnosis could be fatal, especially to the population at risk, such as individuals with compromised immune systems. Intelligent decision systems (IDS) provide a promising tool to expedite discovery of biomarkers, and to boost their impact on earlier prediction of the likelihood of the disease onset. A novel IDS (iTB) is designed that integrates results from molecular medicine and systems biology of Mtb infection to estimate model parameters for prediction of the dynamics of the gene networks in Mtb-infected laboratory animals. The mouse model identifies a number of genes whose expressions could be significantly altered during the TB activation. Among them, a much smaller number of *the most informative* genes for prediction of the onset of TB are selected using a modified version of Empirical Risk Minimization as in Vapnik's statistical learning theory. A hybrid intelligent system is designed to take as input the mRNA abundance at a near genome-size from the individual-to-be-tested, measured 3-4 times. The algorithms determine if that individual is at risk of the onset of the disease based on our current analysis of mRNA data, and to predict the values of the biomarkers for a future period (of up to 60 days for mice; this may differ for humans). An early warning sign allows conducting gene expression analysis during the activation which aims to find key genes that are expressed. With rapid advances in low-cost genome-based diagnosis, this IDS architecture provides a promising platform to advance Personalized Health Care based on sequencing the genome and microarray analysis of samples obtained from individuals at risk. The novelty of the design of iTB lies in the integration of the IDS design principles and the solution of the biological problems hand-in-hand, so as to provide an AI framework for biologically better-targeted personalized prevention/treatment for the high-risk groups. The iTB design applies in more generality, and provides the potential for extension of our AI-approach to personalized-medicine to prevent other public health pandemics.

Keywords: Mycobacterium tuberculosis, biomarkers and intelligent decision systems, early detection, Vapnik's statistical learning theory.

1 Introduction

Tuberculosis is a severe lung disease which is responsible for increase of 9.4 million [1] cases per year that will results about 2 millions of patients death [2]. Infection will be acquired by inhalation of Mycobacterium tuberculosis contaminated air and/or droplets [3]. Even from slight initial invasion of the agent, infection may lead to latent TB or may lead to primary disease [4]. The exact time which is taken from the initial infection until the development of disease varies among individuals. This variation can be attributed mainly to the immune status of an individual. Immuno-compromised individuals or individuals which their immune systems are suppressed might likely develop primary disease. The World Health Organization WHO [1], has reported the burden of disease under various circumstances, and in particular, the number of bacteria sufficient to infect an individual [3]. After 2008, the numbers of incidence, prevalence and mortality, are estimated to be at least as 9.4 million incidence cases, 11.1 million prevalence cases and 1.3 million deaths [1]. Since an individual can infect 10-15 individuals [3], from a public health viewpoint, early diagnosis and treatment [5] is crucial to contain the disease. In particular, technologies for early detection and isolation of an infected individual will play a major role in sustainability of the global population health.

Understanding the molecular mechanisms during the invasion of M. tuberculosis provides valuable insights for the analysis of the biological understanding of the course of infection. Besides, the molecular understanding might lead to development of the necessarily better targeted treatment strategies against tuberculosis [6]. To serve this purpose, Talaat et al. (2004) analyzed MTB infected lung samples of immuno-compromised and immune-competent mice. They found the genes that are expressed (or significantly changed) during early invasion through systematic application of the microarray technology [7]. As a result of this analysis, they reported the differences between expression profiles in three different environments. In an analogous fashion, but in a different context, microarray technology is used to monitor the changes in M. tuberculosis gene expression during the treatment with antituberculous drug isoniazid [8]. According to analysis of expressed genes in presence of isoniazid, researchers are more likely to enhance the drug targets. Behr et al. (1999) perform microarray analysis in order to understand the differences between genomic structures of *M. tuberculosis* and *M. bovis* and other strains of *M. bovis* which are also compound of the BCG vaccines. Accordingly, this study has aimed to serve the purpose of developing new and more narrowly-aimed vaccines and/or antituberculosis treatment [9]. Fisher et al. (2002), draw attention to the function of the acidification during the immune response through using microarray [10]. As a result of their analysis, they suggest that, acidification might be a signal to induce the gene expression needed by the bacteria to survive against the immune response cells known as phagosomes.

The discussion above is short, but essentially highlights the critically sparse state-of-art knowledge regarding detection, prediction and treatment of individuals at risk, and in fact, almost all categories of individuals infected by this microbe.

In the following, we report on preliminary progress in design of an intelligent system (Section 3) based on our earlier *de novo* analysis of gene expression time-series by novel applications of stochastic signal processing, new clustering algorithms, and dynamics of representations of clusters in an appropriate hyperbolic space.

2 Contribution to Sustainability

One remark encountered often in modern higher education and research is the significance of close collaboration between young investigators in the computational and the engineering sciences with biologists to provide a fruitful framework for the synthesis of diverse concepts and tools. In this way, integration of hardware-software and 'biological knowledgeware' provides prospects of imminent solutions of myriad challenging biological sustainability research problems through collaborative effort. This research is an illustration of this piece of educational wisdom in this regard. The need for solution of hard biological problems has inspired formulation of a number of challenging problems in scientific computation and engineering.

High performance scientific computation plays a critical role in the 21st century research and engineering design optimization. An important case arises in research on challenging problems in Global Public Health, such as sustainability of protective measures against infectious diseases for humans and animals living across all geographic locations, and as much as possible, under all states of living conditions. This research adheres to the above-mentioned objective of sustainability, utilizing the state-of-art in informatics and engineering. Design of intelligent systems have enriched the modern technological societies, and extending its domain to include the entire global community is inevitable for future protection of life and earth and assurance of a sustainable mechanism to provide a healthy society across the globe. Further, among myriad health and disease conditions, there is a serious risk of faster spread of diseases such as TB by more susceptible sectors of individuals at risk due to a compromised immune systems. Sustainability of global health, therefore, must include effective solutions to prevent infection and more likely death by such individuals. On the ethical side, there are limitations to keeping in isolation individuals-at-risk due to the higher risks of eventual severe or fatal sickness. The research on iTB system provides a viable approach to answer the above-mentioned problems based on effective applications of modern engineering and informatics.

3 iTB: Intelligent TB Dynamical Modeling

We have developed a conceptual framework for dynamical analysis on the grounds of solid mathematical models and empowered by software engineering viewpoint toward information systems development. This viewpoint ensures reusability of the system as a whole or in part for different applications in different disciplines. In other words, a *modular* design of the system allows us to (1) simplify extending the system capability while maintaining accuracy, (2) rebuild new configurations on

demand - e.g. for different clinical applications, and (3) distribution of massive computation among different processing blocks. The entire setting adheres to the engineering principles for deployment within today's High Performance Computation (HPC) infrastructures, like Computational Grids, and most recently, Computation in Clouds. Data-intensive computing era calls for such adaptable architectures to guarantee the applicability of informatics methods for growing population, increasing healthcare demands and personalized (thus, real-time) medical treatment.

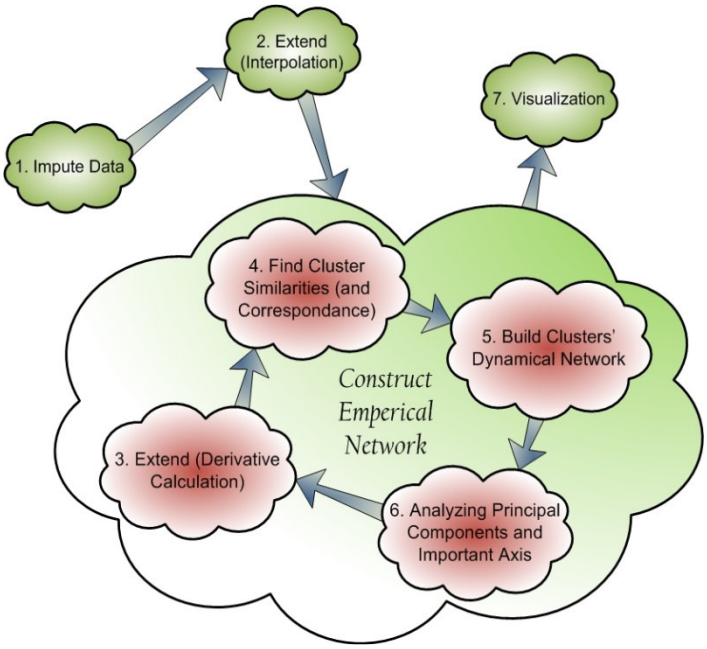

Fig. 1. Global architectural view of iTB

Fig. 1 shows a global architectural view of iTB. The main part is the *empirical network construction* module which can work iteratively to gain computational access into higher levels of dynamics of massively complex biological systems.

A brief description of modules is provided below. The *Imputation Module* tries to fill in the missing pieces of the data, which is a common issue in biological time-series samples. The algorithm utilized here relies on the widely-used Expectation Maximization (EM) algorithm to find the best candidates (i.e. most probable ones) for missing values. We have adapted a generalized version of EM [11] to apply to our problem setting. The next Module also performs preprocessing provides a suitable smoothing of the time-series for mathematical analysis, which requires stable local behavior of the time-series (regarding variations and disregarding spiky results). We have employed piecewise cubic Hermite interpolating polynomial (PCHIP) curve fitting approach [12] and re-sampled the regressed curve to approximately simulate the behavior of the system in an appropriately *smoother* way to accommodate differentiation and other analytic operations.

The next few Modules are designed to find the multi-level structure of the complex networks. They deal with the twisted behavior that results from the connectionism beyond the "*build from the simplest blocks*" philosophy. In non-technical terms, connectionist and other learning-theoretic models are constrained by the nature of the domain of generalization, and the balance between the sufficiency of samples versus overtraining. This implies the requirement of elucidating the interrelationship among samples of different levels of significance for estimating future dynamics from the sample point (among other technical hurdles to overcome the biological complexity in predicting the state of the disease from a sparse sample.) Thus, to capture the dynamics of the disease, hierarchical clustering techniques are employed to build a multi-level structural/behavioral model of interactions inside the system. There are advantages to using hierarchical clustering versus non-hierarchical clustering. A comprehensive analysis of different clustering methods and their applicability will appear in due time. Briefly, hierarchical clustering allows the modeler to take into account various forms of analytic singularities, avoiding the artificial assumption that the data samples are uniformly chosen from a single probability distribution function (pdf). Hierarchical clustering, on the other hand, offers the more realistic assumption that different sub-clusters are samples from several pdf that could somewhat differ from the initial pdf. Such diversity of pdf is expected in biology, due to variation and other complex biological phenomena.

The remaining Modules attempt to render the novel concepts of *special-architecture* empirical networks for topological modeling of complex networks. This approach enables us to use a vast spectrum of solid mathematical analysis tools to reveal invaluable measures of correspondence between components of complicated systems (cf. below for an outline.)

Two of the Modules capture dynamic similarities, and record the migration of different classes of genes with different perspectives towards the inclination of the groups' rate and acceleration change in the course of time. The next Module estimates the probability densities in clusters via the exponential family of models. The exponential family has the unique advantage of being quite flexible to accommodate many deviations from Gaussian models, while still are parameterized via a a finite dimensional Riemannian manifold in the Hilbert space of L^2-integrable functions. The Riemannian structure alluded above is complete and hyperbolic. Hypothesizing the consistent behavior of related genes in the hyperbolic space, we have measured the distance between the clusters of genes as follows, which we mention in the Gaussian case to simplify the presentation.

The individual clusters are regarded as samples from a probability distribution; e.g. for the sake of a concrete illustration, consider two clusters that are regarded as samples from two pdfs that are normal distributions

$$\mu_j \sim N(m_j, \sigma_j^2) \tag{1}$$

So for any time frame, we have a topological representation of the system that evolves in time and shows the behavior of the systems in the frame of the hypothesis. To be able to visually inspect the dynamics of the system, we may argue based on the experimental results that projection of the high-dimensional networks

obtained into the very first principal components gives us a good representation of the behavioral model.

3.1 Mathematical Methods

First, the solution of the preceding estimate requires *transforming the ill-posed problem* into a regularized well-posed problem [13], [14] and [15] Thus, it is desired here to have a well-posed problem regarding *estimating the measure* on the "function-space". There are well-known methods and more modern ongoing research on regularization, and we shall omit the discussion due to lack of space [13] and [15] To readers familiar with learning theory [16], the latter problem could be regarded as "*Learning the Measure"* from the sample of trajectory dynamics (we used the data available to us from the TB-infection of mice) through a controlled iterative scheme. Thus, we proceed to cast the latter in *Statistical Learning Theory*. Accordingly, we need to have a robust estimate for the error in iterative steps of learning to quantify the approximation error for the posterior mentioned above. Robust error estimates require stability in solution of sampling. To have a well-posedness of the inverse problem for the posterior measure will provide desired levels of stability. In turn, such stability may be used as the basis for quantifying the approximation of inverse problems for functions in a finite-dimensional-space setting. This requires an estimate for the distance between the true and approximate posterior distributions, in terms of error estimates for approximation of the underlying forward problem.

Let μ_1 and μ_2 be two normal distributions with means m_1, m_2 and standard deviations σ_1 and σ_2. The computation of distance between them $d_H(\mu_1,\mu_2)$ is as follows

$$\left(1-\sqrt{\frac{2\sigma_1\sigma_2}{\sigma_1^2+\sigma_2^2}}e^{-\frac{1}{4}\frac{(m_1-m_2)^2}{\sigma_1^2+\sigma_2^2}}\right)^{\frac{1}{2}} \tag{2}$$

This metric is initially defined for two measures μ_1 and μ_2 that are absolutely continuous relative to a measure λ as:

$$d_H(\mu_1,\mu_2)=\left(\frac{1}{2}\int\left(\sqrt{\frac{d\mu_1}{d\lambda}}-\sqrt{\frac{d\mu_2}{d\lambda}}\right)^2 d\lambda\right)^{\frac{1}{2}}. \tag{3}$$

where $\frac{d\mu}{d\lambda}$ terms are Radon-Nikodym derivatives respectively. The definition of (3) does not depend on the choice of λ, so (3) does not change if λ is replaced with a different probability measure with respect to which both μ_j are absolutely continuous. For two normal distributions (1), the formula for (3) is (2) and readily could be used, even helped by symbolic algebra, to improve accuracy of the iterative

Learning-theoretic calculations. Control of differences in the Hellinger metric (i.e. d^2_H) leads to control on the differences between expected values of functions and operators (that admit polynomial bounds). Statistical Learning Theory [16] provides the tools to complete the remaining steps in this approach. There are other computational reasons for choice of the Hellinger metric versus other probability-theoretic divergences, say from the family of f-divergences such as the Kullback-Leibler, Wasserstein or other metrics. Among them, one could gain useful estimates of bounds more easily using this metric, such as in

$$\| \int f d\mu_1 - \int f d\mu_2 \| \leq 2\left(\int \| f \|^2 \, d\mu_1 - \int \| f \|^2 \, d\mu_2\right)^{\frac{1}{2}} d_H(\mu_1, \mu_2) \quad \textbf{(5)}$$

In turn, such bounds point to design of numerical schemes that allow us to solve the inverse problem and gain control in relating estimates arising during perturbations in the domain and range, respectively. Briefly, in the discussion above, let the integer N denote the iteration count, and correspondingly, the estimates $\Psi^N(v;w)$, μ^N, such that $\frac{d\mu}{d\mu_0} \sim e^{-\Psi(v;w)}$, $\frac{d\mu^N}{d\mu_0} \sim e^{-\Psi^N(v;w)}$. Then bounds on (6) provides a sequence of improving bounds on (3) that demonstrates when $N \to \infty$, the Hellinger metric approaches zero exponentially in N.

$$| \Psi(v\,;w) - \Psi^N(v\,;w) | \quad \textbf{(6)}$$

For Gaussian densities, these bounds are used to prove that the means and covariance operators associated to μ^N, and μ, *are* close in the Hilbert-space (or the Banach space, in more general circumstances) operator-norms. Therefore, in the approach outlined above, we could arrange for the transfer of estimates from the numerical analysis of forward problems into estimates for the solution of the related inverse problem.

4 Results and Discussion

In the preceding arguments, the space of exponential probability density functions is a Riemannian manifold, and we need a discrete approximation to capture its metric properties within the prescribed error bound. The discrete approximation is typically expected to be high dimensional (thousands or more). In the case of normal distributions, a number of analytic simplifications are available that allow of us to reduce the dimensionality of the metric model. In particular, the Riemannian metric could be approximated by sampling of the distance data, and the model reduction for the sample agrees with the desired approximation to the Riemannian structure. As expected Singular Value Decomposition provides the direct approach, and in the case of animal models of the disease (murine), the results are obtained as follows. Empirically, we have observed that considering the first three principal components retains about 80-85% of the information content of the network (i.e. the ratio of the first 3 eigenvalues to the total sum), while the dynamical separation of different conditions under study are brilliantly visible. Figures 2 to 7 show a sample dynamics

of the TB genes in the level of data. The axes are all relative to a unit-free representation of the cluster distances as measured in the Heilinger distance. Figures are converted to the 3-dimensional projection metric for visualization purposes. The figures are shown in different zoom levels for clarity purposes.

We have studied, as an example, the dynamics of gene expression profiles of TB-related families in a period of 24 hours after infection. Profiles are recorded every 4 hours and a sliding window approach is used to investigate the differences between the "behavioral associations" among genes. As could be observed in the figures, there is a heart-beat-like pattern between clusters representing the conditions. Each point on the graphs represents a cluster of genes in one of the conditions. The observable dynamics resemble a group of particles in a force field which approach each other and then the repulsing forces cause the system to scatter around.

The importance of the above-mentioned results is that the behavioral differences between the two conditions are observable in the very first stages of getting exposed to the invader. The discrimination between the two groups of points (red crosses vs. blue circles) which correspond to the different conditions, could be observed from these graphs as a growth/shrinkage pattern. From a clinical treatment viewpoint, this is of utmost importance as one could be treated before the infection spreads out of control.

At the time of writing this article, we are making progress in translating the above-mentioned visualization-based observations (the differences between the two cases) into a classification scheme within the 3-dimensional principal component space (please see figures). Our method is based on design of new algorithms that minimizes the empirical risk function (analogous to the Vapnik soft-margin SVM empirical risk functional), using the hyperbolic metric in lieu of the Euclidean distance in the theory by Vapnik and others. While the computations are much more challenging, the mathematical results that must guarantee the existence of a minimum for the risk functional and the desired regularity properties are ensured through extensive mathematical work on analysis of functions on hyperbolic spaces, in particular, bounded sequences of approximations to a minimum converge exponentially (hence even faster than the Euclidean metric) and there is a unique limit point for the sequence, hence a unique minimum. The details will be provided in a forthcoming paper.

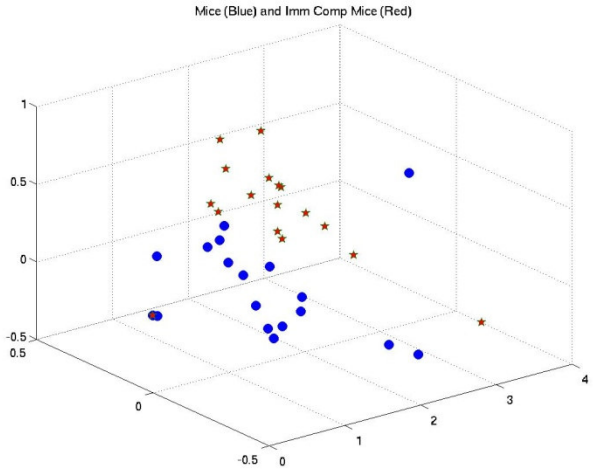

Fig. 2-7. These samples of the visualization movie frames show instances of patterns in dynamics of the hyperbolic representation of the analyzed sample of TB gene expressions in time. The dynamics is approximated discretely, then projected to the 3-dimensional reduced model from the hyperbolic space. The first three dominant principal modes capture more than 80% of the information contents in the original hyperbolic space. Once the separations of different dynamic patterns are accomplished in the reduced model, clearly the original data will also demonstrate the separation by considering the inverse-images of the separating hyperplanes.

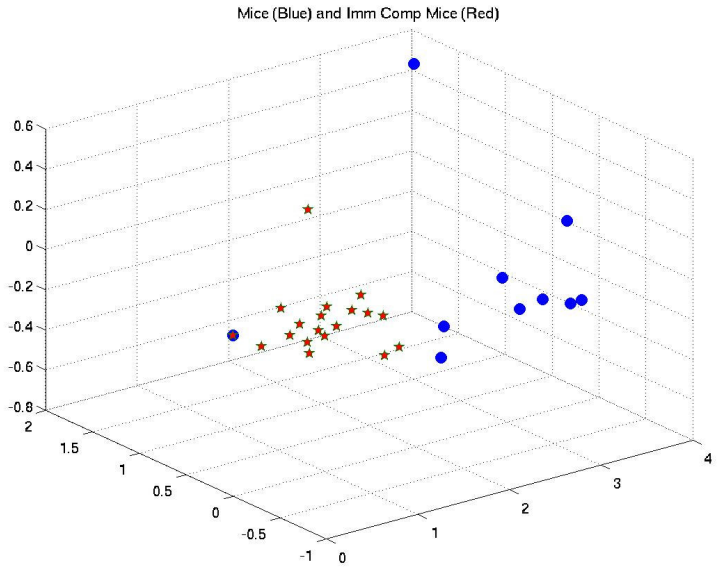

Fig. 3 A sample of projection of the gene expression dynamic pattern in the reduced model.

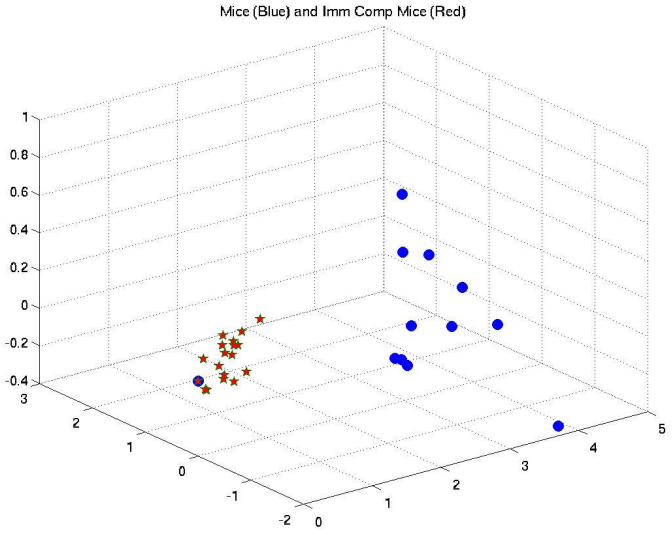

Fig. 4 A sample of projection of the gene expression dynamic pattern in the reduced model.

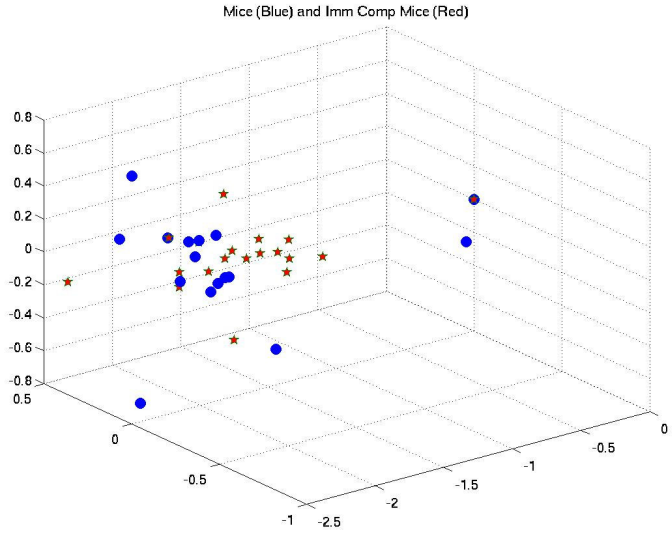

Fig. 5 Another sample of projection of the gene expression dynamic pattern in the reduced model

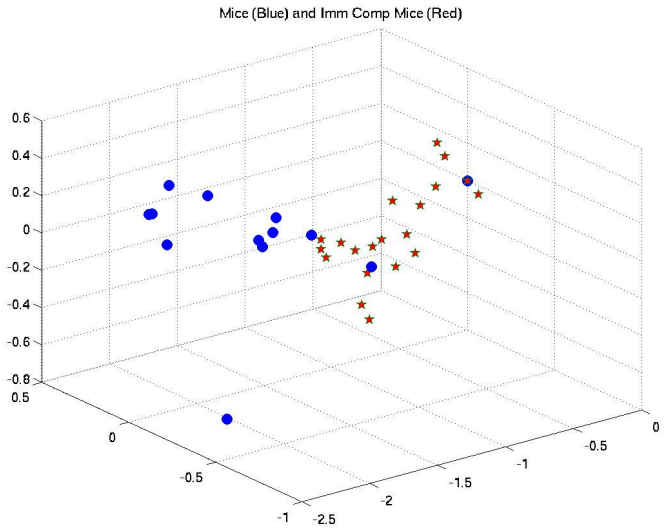

Fig. 6 Another sample of projection of the gene expression dynamic pattern in the reduced model.

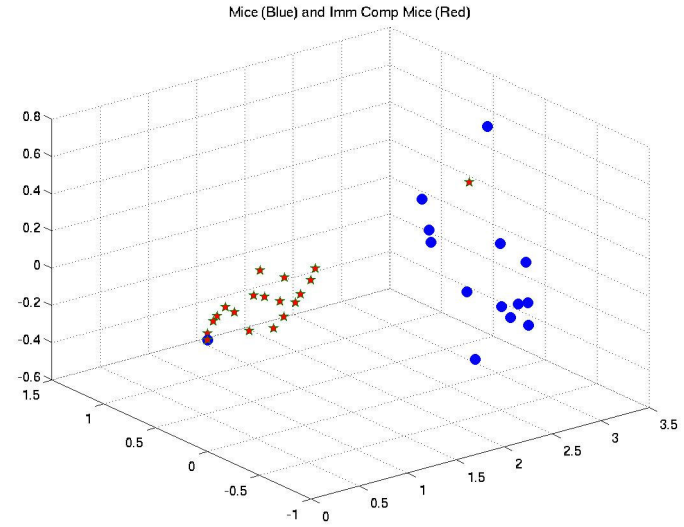

Fig. 7 Another sample of projection of the gene expression dynamic pattern in the reduced model.

References

1. World Health Organization: Global Tuberculosis Control: A short update to the 2009 report.
2. Volokhov, D.V., Chizhikov, V.E., Denkin, S., Zhang, Y.: Mycobacteria Protocols. Humana Press, New York (2008).
3. World Health Organization, http://www.who.int/topics/tuberculosis/en/
4. Murray, M.: Tuberculosis: The Essentials. Informa Healthcare, New York (2010)
5. Hopewell, P.C.: Tuberculosis: The Essentials. Informa Healthcare, New York (2010)
6. Triccas,J.A., Berthet,F.X., Pelicic,V., Gicquel,B.: Use of Fluorescence Induction and Sucrose Counter Selection to Identify Mycobacterium tuberculosis Genes Expressed Within Host Cells. Microbiology 145, 2923--2930 (1999)
7. Talaat, A.M., Lyons, R., Howard, S.T., Johnston S.A.: The Temporal Expression Profile of Mycobacterium tuberculosis Infection in Mice. Proc.Natl.Acad.Sci, pp. 4602--4607. PNAS, USA (2004)
8. Wilson, M., DeRisi, J., Kristensen, H.H., Imboden, P., Rane, S., Brown,P.O., Schoolnik, G.K.: Exploring Drug-induced Alterations in Gene Expression in Mycobacterium tuberculosis by Microarray Hybridization. Proc.Natl.Acad.Sci. pp. 12833--12838. PNAS, USA (1999)
9. Behr, M.A., Wilson, M.A., Gill, W.P., Salamon, H., Schoolnik, G.K., Rane, S., Small, P.M.: Comparative Genomics of BCG Vaccines by Whole-Genome DNA Microarray. Science 284, 1520--1523 (1999)
10. Fisher, M.A., Plikaytis, B.B., Shinnick, T.M.: Microarray Analysis of the Mycobacterium tuberculosis Transcriptional Response to the Acidic Conditions Found in Phagosomes. J.Bacteriol. 184, 4025--4032 (2002)
11. Schneider, X.: Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values. Journal of Climate. 14, 853--871 (2001)
12. Fritsch, F. N., Carlson,R. E.: Monotone Piecewise Cubic Interpolation. SIAM J. Numerical Analysis. 17, 238--246 (1980)
13. Poggio, T., Girosi, F.: Regularization algorithms for learning that are equivalent to multilayer networks. Science. 247: 978--982 (1990)
14. Girosi, F.: An Equivalence Between Sparse Approximation and Support Vector Machines. Neural Computation. 10, 1455--1480 (1998)
15. Smola, A.J., Schölkopf, B.: Form Regularization Operators to Support Vector Kernels. Morgan Kaufmann (1998)
16. Vapnik, V.N.: The Nature of Statistical Learning Theory. Springer, New York (2000)