A Survey of Transfer and Multitask Learning in Bioinformatics

Qian Xu and Qiang Yang*

Hong Kong University of Science and Technology, Kowloon, Hong Kong, China
fleurxq@ust.hk, qyang@cse.ust.hk

Abstract

Machine learning and data mining have found many applications in biological domains, where we look to build predictive models based on labeled training data. However, in practice, high quality labeled data is scarce, and to label new data incurs high costs. Transfer and multitask learning offer an attractive alternative, by allowing useful knowledge to be extracted and transferred from data in auxiliary domains helps counter the lack of data problem in the target domain. In this article, we survey recent advances in transfer and multitask learning for bioinformatics applications. In particular, we survey several key bioinformatics application areas, including sequence classification, gene expression data analysis, biological network reconstruction and biomedical applications.

I. INTRODUCTION

With the fast growth of biological technology, there has been a rapid growth in the need to process biological data. This data has been made available with lower costs via advanced bio-sensor technologies, and as a result, in the next few years, we expect to witness a dramatic increase in the application of genomics in our everyday lives, such as in our personalized medicine.

Bioinformatics research is interdisciplinary in nature, in that it integrates diverse areas such as biology and biochemistry, data mining and machine learning, database and information retrieval, computer science theory, as well as many others. Bioinformatics concerns a wide spectrum of biological research, including genomics and proteomics, evolutionary and systems biology, etc. These research areas are increasingly dealing with data collected from a wide spectrum of devices such as microarrays, genomic sequencing, medical imaging and so on [1]. Based on this data, data mining in bioinformatics targets the investigation of learning statistical models to infer biological properties from new data. Various techniques such as supervised learning and unsupervised learning have been developed, with promising results and biological insights in areas such as sequence classification, gene expression data analysis and biological network reconstruction, etc.

A typical assumption in biological data mining is that a sufficient amount of annotated training data is required, so that a robust classifier model can be trained. In many real-world biological scenarios, however, labeled training data is lacking. Sometimes this data can only be obtained by paying a huge cost. This problem of a lack of labeled training data, which partly causes the so-called data sparsity problem, has become a major bottleneck in practice. When data sparsity occurs, over fitting is common when we train a model. As a result, the statistical model will experience a reduction in performance.

In response to the above problem of data sparsity, various novel machine-learning methods have been developed. Among them is the transfer learning framework, which refers to a new machine learning framework which reuses knowledge from other domains. Transfer learning aims to extract knowledge from some auxiliary domains where the data is annotated, but it cannot be directly used as the training data for a new domain. To reuse this data, transfer learning compares auxiliary and target problem domains in order to find useful knowledge that is

common between them, and then reuses this knowledge to help boost learning performance in the target domain. A subarea of transfer learning is multi-task learning, where the target task and the source task are treated the same, and are learned simultaneously. Multitask learning stresses generating benefits for learning performance improvements in all related domains. Transfer learning and multi-task learning have achieved great success in many different application areas over the last two decades [2, 3], including text mining [4-6], speech recognition [7, 8], computer vision (e.g., image [9] and video [10] analysis), and ubiquitous computing [11, 12].

In this article, we give a survey on some recent transfer learning methods in bioinformatics research, together with an additional overview on biomedical imaging and sensor based e-health. We will cover some of the most important application areas of bioinformatics, including sequence analysis, gene expression data analysis, genetic analysis, systems biology and biomedical applications in bioinformatics. By conducting a survey in these areas, we hope to take stock of the progress made so far, and help readers navigate through technological advances in the future.

II. DATA MINING IN BIOINFORMATICS: A SURVEY OF PROBLEMS

We focus on the following biological problems in this survey: sequence analysis, gene expression data analysis and genetic analysis, systems biology, biomedical applications.

A. Biological Sequence Analysis

Biological sequence analysis aims to assign functional annotations to sequences of DNA segments, and is important in our understanding of a genome. One example is the identification of splice sites in terms of the exon and intron boundaries, a complex task due to the number of alternative splicing possible. Other examples include the prediction of regulatory regions that allow the binding of proteins and determine their functions; the prediction of transcription start and initiation sites, and the prediction of coding regions. Another important sequence analysis problem is the major histocompatibility complex (MHC) binding prediction from a sequence perspective, where sequences from pathogens provide a huge amount of potential vaccine candidates [13]. MHC molecules are key players in the human immune system, and the prediction of their binding peptides helps in the design of peptide-based vaccines, but there is a major lack of quality labeled training data that can be used to build a good prediction model for this binding prediction problems. In addition, protein subcellular localization prediction problems are also important in biological sequence analysis, but annotated locations are difficult to obtain.

B. Gene Expression Analysis and Genetic Analysis

Gene expression analysis and genetic analysis through microarrays or gene chips is an important task for the understanding of proteins and mRNAs. A microarray experiment measures the relative mRNA levels of genes, which allows us to compare the gene expression levels of some biological samples over time in order to understand the differences between normal cells and cancer cells [14]. One characteristic of this analysis is that the number of features that correspond to genes is usually more than the number of samples. This makes it difficult to apply traditional feature selection approaches directly to this data to reduce its dimensionality [15]. Another data mining task often applied to gene expression data is co-clustering, which aims to cluster both the samples and genes at the same time [16]. In genetic analysis, recent advances have allowed genome-wide association studies (GWAS) that can assay hundreds of thousands of single nucleotide polymorphisms (SNPs) and relate them to clinical conditions or measurable traits. A combination of statistical and machine learning methods has been exploited in this area [17, 18].

C. Computational Systems Biology

Computational systems biology refers to the tasks of modeling gene-protein regulatory networks, and making inferences based on protein-protein interaction networks. The computational challenge relates to the task of integrating and exploring large-scale, multi- dimensional data with multiple types. An important issue is how to exploit the topological features of the network data to help build a high-quality model. Besides automatic prediction based on statistical models, mixed initiative approaches such as visualization have also been developed to tackle the large-scale data and complex modeling problems.

D. Biomedical Applications

Biomedical applications investigated in this survey include biological text mining, biomedical image classification and ubiquitous healthcare. Biological text mining refers to the task of using information retrieval techniques to extract information on genes, proteins and their functional relationships from scientific literature [19]. Today we face a vast amount of biological information and findings that are published as articles, journals, blogs, books and conference proceedings. PubMed and MEDLINE provide some of the most up to date information for biological researchers. A scientist often has to read through a large volume of published text in order to understand and find any potential knowledge embedded in the text. With text mining technology, many new findings that are published in text format, such as new genes and their relationships, can be automatically detected, which helps man and machine work together as a team.

Furthermore, biomedical image classification is an important problem that appears in many medical applications. Manual classification of images is time-consuming, repetitive, and unreliable. Given a set of training images labeled into a finite number of classes, our goal is to design an automatic image classification method to understand future images. An example of this is breast-cancer identification from medical imaging data using computer-aided detection on the screening mammography. An important issue for such models is how to reduce the false-positive rates in the classifications.

Ubiquitous computing is increasingly influencing health care and medicine and aims to improve traditional health care [20].

Exercises are essential to keep people healthy. Recently, several devices have been designed to record and monitor users' routine exercises. For example, the accelerometer [21-23] and gyroscope [24-26] are used to recognize the users' activities; Polar records heart rate to evaluate the amount of activity in terms of calories in order to manage sports training [27]. Pedometers can quantify activities in terms of number of steps, and then suggest the users to reduce or increase the amount of exercise they undertake [28].

In all the above-mentioned problems, transfer learning plays an important role for solving the data sparsity problem. Below, we survey some recent transfer learning studies in these areas. To start with, we first introduce some notations first that are used in this survey. The source domain data D_S is composed of data instances $x_i \in X_S$ with their corresponding labels $y_i \in Y_S$. Thus we denote the source domain data (X^S, y^S) as $\{(x_1^S, y_1^S),..., (x_{n_S}^S, y_{n_S}^S)\}$. Similarly, the target domain data D_T is composed of data instances $x_i \in X_T$ with their corresponding labels $y_i \in Y_T$. We denote the target domain data (X^T, y^T) as $\{(x_1^T, y_1^T),..., (x_{n_T}^T, y_{n_T}^T)\}$. The functions $f_S(\cdot)$ and $f_T(\cdot)$ denote predictive functions in the source domain D_S and the target domain D_T, respectively, where $y^S ; f_S(X^S)$ and $y^T ; f_T(X^T)$. When considering multi-task situations, the data (X^t, y^t) of each task $t \in TN$ can be represented by $\{(x_1^t, y_1^t),..., (x_{n_t}^t, y_{n_t}^t)\}$, where TN is the number of tasks for multi-task learning.

III. BIOLOGICAL SEQUENCE ANALYSIS

In sequence classification, the goal is to annotate gene sequences or protein sequences from a given set of training data. As we mentioned above, the process of learning to annotate sequences often suffers from a lack of labeled data, which often leads too verfitting. To solve this problem, multi-task learning methods are often used, where it possible to learn to annotate two or more sets of sequence data together. In this approach, the sequence data can be from different problem domains. By learning these tasks together, the lack of labeled data problem is alleviated.

In multi-task learning, task regularization is an often used. Task regularization methods are formulated under a regularization framework, which consist of two objective function terms: an empirical loss term on the training data of all tasks, and a regularization term that encodes the relationships between tasks. We first introduce a general framework of regularization based multi-task approach. In their pioneering work, Evgeniou and Pontil [29] proposed a multi-task extension of support vector machines (SVMs), which minimizes the following objective function,

$$\xi(\{w_t\}) = \sum_{t=1}^{TN}\sum_{i=1}^{n_t} l(y_i^t, w_t^T x_i^t) + \lambda_1 \sum_{t=1}^{TN}\|w_t\|_2^2$$
$$+ \lambda_2 \sum_{t=1}^{TN}\left\|w_t - \frac{1}{TN}\sum_{t=1}^{TN}w_t\right\|_2^2 \qquad (1)$$

The first and second terms of Eq. (1) denote the empirical error and 2-norm of parameter vectors, respectively, which are the same as those in single-task SVMs. In biological sequence classification tasks, these terms refer to the errors committed when making a comparison to the training data. The difference between single-task SVMs and multi-task SVMs lies in the third term of Eq. (1), which is designed to penalize large deviations between each parameter vector and the mean parameter vector of all tasks. The penalized term imposes a constraint that the parameter vectors in all tasks be similar to each other. In multi-task sequence classification problems, this term characterizes the extent of knowledge sharing between the learning tasks in different sequences. With this general background on SVM-based multitask learning in mind, we now sample a few well-known works using regularization based approaches, as well as kernel design for biological sequence analysis. We will pay attention to applications in splice site recognition and MHC binding prediction.

One of the first works in multi-task biological sequence classification was by Widmer et al. [30], who proposed two regularization based multi-task learning methods to predict the splice sites across different organisms. In order to leverage information from related organisms, Widmer et al. [30] suggested two principled approaches to incorporate relations across various organisms. The proposed methods were implemented by modifying the regularization term in [29]. However, the relations between the organisms in [30] was defined by a tree or graph implied by their taxonomy or phylogeny. The models related to task t served as prior information for the model of t, such that the parameter w_t of task t is required to be close to the parameters w_{others} of the other models. This can be achieved by minimizing the norm of the differences of the parameter vectors, $\|w_t - w_{others}\|$, along with the original loss function. The first approach trains the models in a top-down manner, where a model learns for each node in the hierarchy over the datasets of the tasks spanned by the node and the parent nodes which are used as the prior information, $\|w_t - w_{parent(t)}\|$. The objective function was given as follows:

$$\xi(\{w_t\}) = \sum_{t=1}^{TN}\sum_{i=1}^{n_t} l(y_i^t, w_t^T x_i^t) + \lambda_1 \sum_{t=1}^{TN}\|w_t - w_{parent(t)}\|_2^2 \qquad (2)$$

In biology, an organism and its ancestors should be similar due to inheritance of properties in evolution. This is imposed as a constraint in the above formula. Another approach adopted by Widmer et al. [30], is the constraint that $\|w_t - w_{t'}\|$ must be small for related tasks t and t' in their formulation.

$$\xi(\{w_t\}) = \sum_{t=1}^{TN}\sum_{i=1}^{n_t} l(y_i^t, w_t^T x_i^t) + \lambda_1 \sum_{t=1}^{TN}\sum_{t'=1}^{TN}\gamma_{tt'}\|w_t - w_{t'}\|_2^2 \qquad (3)$$

Intuitively, the above formulation states that the properties associated with similar entities are close to each other. $\gamma_{tt'}$ reflects evolutionary history before the divergence of the two organisms.

In addition to the above formulations, Widmer et al. [30] also designed a kernel function that not only considers the data of the task t, but also the data of its corresponding ancestor r_t, according to a hierarchical structure R, which is defined manually. The kernel can be derived by defining prediction functions over the task parameters as well as the parameters of the ancestors in terms of the tasks.

$$\xi(\{w_t\}) = \sum_{t=1}^{TN}\sum_{i=1}^{n_t} l(y_i^t, w_t^T x_i^t) + \lambda_1 \sum_{t=1}^{TN}\|u_t\|_2^2 + \lambda_2 \sum_{t=1}^{TN}\sum_{r_t=1}^{R}\|v_{r_t}\|_2^2 \qquad (4)$$
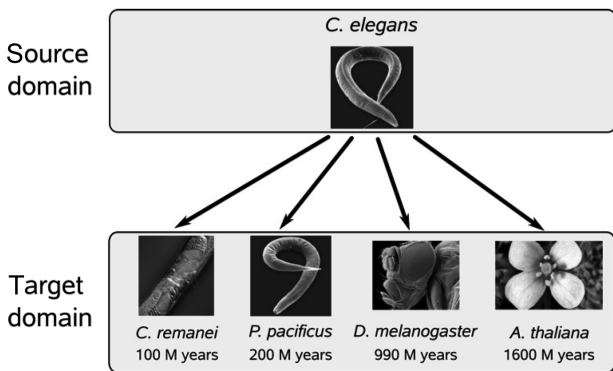
Fig. 1. Illustrating source and target domains in biology (adopted from invited talk by Gunnar Rätsch, Invited Talk at NIPS Transfer Learning Workshop, December 2009, Whistler, B.C. (Schweikert et al. [31]).

Fig. 2. Four domain adaptation models (adopted from invited talk by Gunnar Rätsch, Invited Talk at NIPS Transfer Learning Workshop, December 2009, Whistler, B.C. (Schweikert et al. [31]).
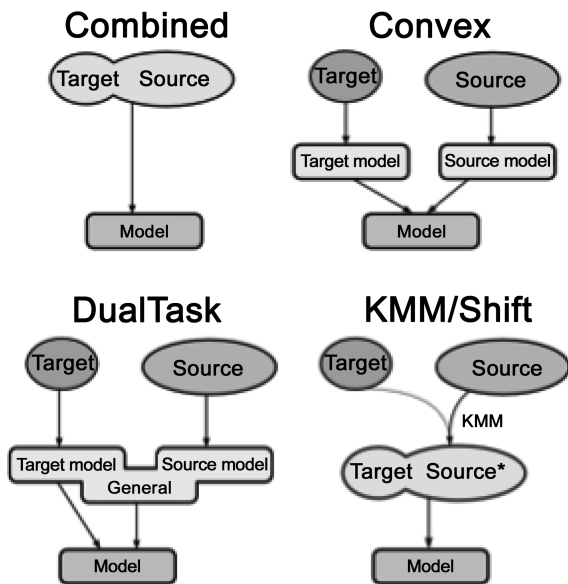
In Eq. (4), u represents the parameter vectors of leaf nodes and v represents the parameter vectors of their corresponding ancestors, which are internal nodes in the hierarchy structure.

Schweikert et al. [31] considered a number of domain transfer learning methods for splice-site recognition across several organisms. The goal of domain transfer learning is to use a model from a well-analyzed source domain and its associated data to help train a model in the target domain. In the formulation of Schweikert et al. [31], *C. elegans* is the source domain and *C. remanei*, *P. pacificus* and so on are target domains (Fig. 1).

The domain transfer learning methods include (as illustrated in Fig. 2):
- Combination: As a baseline, the simplest way is to combine the source domain data and the targetdomain data directly without considering the weights.
- Convex combination:

$$F(x) = \alpha * f_T(x) + (1 - \alpha) * f_S(x) \qquad (5)$$

where α is the trade-off parameter, which is used to balance the contributions of the source data and the target data.
- Dual-task learning:

$$\xi(\{w_S, w_T\}) = C \sum_{i=1}^{n_S + n_T} l(y_i^{S+T}, w^T x_i^{S+T}) + \lambda \|w_S - w_T\| \qquad (6)$$

where both of w_S and w_T are optimized and can be solved using a standard QP-solver.
- Kernel mean matching:

$$\hat{\Phi}(x_k) = \Phi(x_k) - \alpha \left(\frac{1}{n_S} \sum_{i=1}^{n_S} \Phi(x_i) - \frac{1}{n_T} \sum_{i=n_S+1}^{n_S+n_T} \Phi(x_i) \right) \forall_i = 1,...,n_S \qquad (7)$$

$$K(x_i, x_t) = \langle \Phi(x_i), \Phi(x_t) \rangle$$

where Φ is the kernel mapping, which projects the data into a reproducing kernel Hilbert space.

Experimental results presented in [31] showed that the differences of classification functions for recognizing splice site in these organisms will increase with increasing evolutionary distance.

Jacob and Vert [32] designed an SVM algorithm that is able to learn peptide-MHC-I binding models for many alleles simultaneously, by sharing binding information across alleles. The sharing of information is controlled by a user-defined measure of similarity between alleles, where the similarity can be defined in terms of supertypes, or more directly by comparing key residues known to play a role in the peptide-MHC binding. Jacob and Vert first represented each pair of alleles a and peptide candidates p by a feature vector. Then, using the kernel trick, they gave the inner product of pairs of alleles and peptides as,

$$K((p, a), (p', a')) = K_{pep}(p, p') \times K_{all}(a, a') \qquad (8)$$

where for the peptide kernel K_{pep}, any existing kernel between peptide representations can be plugged in. For the allele kernel K_{all}, the authors exploited some inner product computation methods to model the relationships across alleles. These allele kernels include the multitask kernel and supertype kernel [32]. In this formulation, alleles have corresponding multitask kernels such that the training peptides that are shared under the same supertype and are trained together, allowing the kernels to reflect the relative distances between the alleles.

In the work of Jacob et al. [33], regularization based models are used for MHC class-I binding prediction. A novel, multitask learning method was implemented by designing a norm or a penalty term over the set of weights, such that the weight vectors of the tasks within a group are similar to each other. To achieve this goal, the penalty function was formulated as follows:

$$\Omega(W) = \varepsilon_M \Omega_{mean}(W) + \varepsilon_B \Omega_{between}(W) + \varepsilon_W \Omega_{within}(W) \qquad (9)$$

where $\Omega_{mean}(W)$ measures the average of weight vectors, $\Omega_{between}(W)$ is a measure of inter-cluster variance and $\Omega_{within}(W)$ is a measure of in-cluster variance.

Following the pioneering works of Jacob and Vert [32], Jacob et al. [33], and Widmer et al. [34] in 2010 aimed to improve the predictive power of the multitask kernel method for MHC class I binding prediction by developing an advanced kernel based on Jacob and Vert's orginal. In addition, Widmer et al.

[35] investigated multi-task learning scenarios where there exists a latent structure shared across tasks. They modeled the crossover between tasks by defining meta-tasks, so that information is transferred between two tasks t and t' with respect to their relatedness and according to the number of meta-tasks in which t and t' co-occur. The importance of the meta-tasks was defined by the learned mixture weights.

As mentioned in Section II, protein subcellular localization prediction can be categorized as biological sequence analysis. Xu et al. [36] compared a multitask learning method with a common feature representation. A common feature representation approach was adapted from Argyriou et al. [37, 38]. In the latter paper, Argyriou et al. [38] proposed a multi-task feature learning method to learn common representations for multi-task learning under a regularization framework:

$$\xi(U, A) = \sum_{t=1}^{TN} \sum_{i=1}^{n_t} l(y_i^t,\ \alpha_t^T U^T x_i^t) + \lambda \|A\|_{2,1}^2 \qquad (10)$$

In the above formulation, the first term $l(\cdot, \cdot)$ denotes the loss function. U is the common transformation to find a common representation. The second term is a regularization term that penalizes the (2,1)-norm of the matrix A, this aims to enforce a scarcity of common features across the tasks. More specifically, $\|A\|_{2,1}^2$ first computes $\|\alpha^i\|_2$, the 2-norms of the rows of matrix A and then computes the 1-norm of the vector ($\|\alpha^1\|_2, ..., \|\alpha^{n_1}\|_2$). This favors solutions in which entire rows of A are zero, which encourages selecting only features that are useful for all tasks.

Xu et al. [36] conducted extensive experiments to answer the question: "Can multi-task learning generate more accurate classifiers than single task learning?" They compared the accuracies of the test data between the proposed multi-task learning methods and other prediction model baselines. An example of their results is summarized in Fig. 3, this illustrates the accuracies of multi-task feature learning and those of single task learning. The columns from left to right and rows from up to down represent different organisms: archaea, bacteria, gneg, gpos, bovine, dog, fish, fly, frog, human0, human, mouse, pig, rabbit, rat, fungi, plant0, plant, virus0 and virus in order. Each cell C_{ij} in the figure is the average result found over 5 random trails. More specifically, C_{ij} is the result of jointly training these models on the organism i and the organism j, and using the trained model $f_j(\cdot)$ with the test data from the organism j. For diagonal cells C_{ij} (shown in gray), Xu et al. [36] trained models on the training data of organism i only, and evaluated the test data of organism

i as well; these were used as baselines. The cells marked in red indicate that applying multi-task learning methods gives worse performance than the baselines, whereas those in light green indicate that applying multi-task learning methods achieves better performance than the baselines. Furthermore, the cells in dark green or dark grey represent the best performance when evaluating the test data from each organism in a particular column. Finally, the cells in white mean the performance result is missing because some organisms overlap, as in the case of human0 vs. human, plant0 vs. plant and virus0 vs. virus. For these pairs, one cannot conduct multi-task learning experiments.

As shown in above figures, the most significant improvement when using the multi-task learning strategy was about 25%. The performance of plants, viruses and animals can be improved by around 10% by using multi-task learning methods. The columns from left to right and rows from up to down in the table show different organisms: archaea, bacteria, gneg, gpos, bovine, dog, fish, fly, frog, human0, human, mouse, pig, rabbit, rat, fungi, plant0, plant, virus 0 and virus in order. This order arranges learning tasks based on organisms in super type order; for example, those organisms that are animals were put together. Moreover, better results are often obtained near the diagonals, where organisms are similar, while worse cases were often located in the cells far from the diagonals. The results in cells near the diagonals were obtained by training using two relatively similar tasks, like dog and fly, bacteria and archaea, and so on. To conclude, multi-task learning techniques can generally help improve the prediction performance for protein subcellular localization in comparison with supervised single-task learning techniques. Furthermore, the relatedness of tasks may affect the final performance in the multi-task learning framework.

Liu et al. [39] proposed a cross-platform model based on multi-task linear regression. They applied the model to multiple datasets for small interfering RNA (siRNA) efficacy prediction. Given a representation of siRNAs as feature vectors, a linear ridge regression model was applied to predict novel siRNA efficacy from a set of siRNAs with known efficacy. Experiments were run to compare the proposed multi-task learning and traditional single task learning based on root mean squared error. It was shown that in siRNA efficacy prediction, there exists a certain efficacy distribution diversity across the siRNAs bindings onto different mRNAs. Common properties across different siRNAs have influence on potent siRNA design. One way to represent the gene expression data is via a matrix, where gene samples are rows and gene expression conditions are columns. There are two categories of rows: control or case, for any sample. The objective of gene expression classification is to accurately classify new samples based on conditions into case or control. This problem is particularly challenging because the data is very imbalanced: the number of samples are often smaller than the number of features (columns). Furthermore, the data can be very noisy.

Chen et al. [40] proposed a multi-task method, known as support vector sample learning (MTSVSL), and demonstrated that the genes selected by MTSVSL yield superior classification performance when using cancer gene expression data. They started by extracting significant samples that reside on support
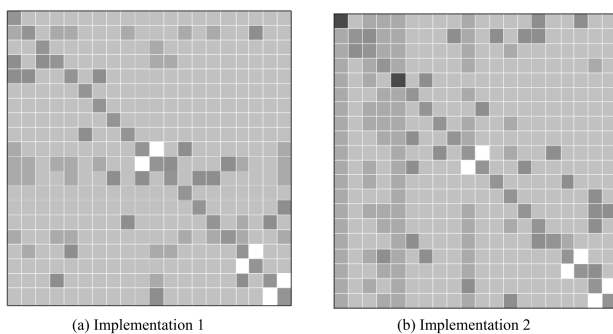

(a) Implementation 1   (b) Implementation 2

Fig. 3. Summary of determined performances.

vectors, and then learned two tasks using multiple back-propagation neural networks simultaneously. They configured their system so that the output of the second task can help refine the original model. One task is aimed at answering "what kind of sample is this," and the second task is "is this sample a support vector sample?" Their experimental results show that the second task can improve classifier performance by incorporating the generated bias with the original bias obtained from the first task.

In the area of genetic analysis, one of the main issues faced today is GWAS. In this area, Puniyan et al. [41] developed a novel multitask regression based technique to perform joint GWAS mapping on individuals from multiple populations. Initially assuming that there is no population structure, the a lasso function can be formulated as:

$$\xi(\beta^t) = \frac{1}{2}(y^t - X^t\beta^t)'g(y^t - X^t\beta^t) + \lambda\|\beta^t\|_1 \qquad (11)$$

where β^t denotes the in-population association strength and $t \in P$ denotes the population index. A limitation of the method is that the model analyzes each population separately without taking advantage of the relatedness among the shared causal SNPs, which may lead it to miss the weak association signals for common SNPs. Puniyani et al. [41] subsequently introduced L1/L2-regularizer for detecting SNPs in different data,

$$\xi(B) = \frac{1}{2}(y^t - X^t\beta^t)'g(y^t - X^t\beta^t) + \lambda\|B\|_{L_1/L_2} \qquad (12)$$

In the above equation, B is a $m \times P$ matrix, m is the number of SNPs genotyped and the j-th row β_j corresponds to the j-th SNP in population P. Here, L_1 / L_2 ensures that the coefficients β_j for the j-th SNP are minimized across all populations. This helps by reducing the number of false positives.

Both gene expression data analysis and genetic analysis suffer from a lack of sample data due to the high cost of data collection in biology. However, the existence of gene expression data under various platforms or conditions, and genome-wide SNP data of different populations provides us with an opportunity to explore relatedness among multiple datasets to alleviate this data scarcity problem. So far, there have only been a few works in this field [40-42] that use transfer learning methods in machine learning. Future studies are expected to explore how to find the most informative genes or SNPs when classifying different sample conditions by investigating their global characteristics among multiple datasets. This objective can be achieved by regularization-based approaches, learning common feature representation approaches, and distribution matching approaches.

IV. SYSTEMS BIOLOGY

In recent years, the application of transfer learning to systems biology has become increasingly popular. Transfer learning techniques implemented by task regularization, distribution matching, matrix factorization and Bayesian approaches have all been employed in computational systems biology.

Gene interaction network analysis has been very useful for gaining insights into various cellular properties. In 2005, Tamada et al. [43] utilized evolutionary information of two organisms to reconstruct their individual gene networks. Suppose that we have two organisms A and B with respective gene expression data D_A and D_B. The networks of the two organisms G_A and G_B are built simultaneously by a hill-climbing algorithm that maximizes the posterior probability function $P(G_A,G_B|D_A,D_B,H_{AB})$ where H_{AB} models the evolutionary information shared between A and B. In order to calculate $P(H_{AB}|G_A,G_B)$ based on the gene expression data D_A and D_B, two free parameters need to be set. In their approach, these two parameters are chosen empirically.

As a follow-up, in 2008, Nassar et al. [44] proposed a new score function that captures the evolutionary information shared between A and B by a single parameter β, instead of choosing two free parameters. The inputs of their multitask learning algorithm now include data samples D of the given organism, an input directed acyclic graph (DAG) G_{in} of the other organism, and a similarity parameter β. The parameter β in their work represents the similarity of the true underlying Bayesian networks. The output derived from the learning algorithm is an improved DAG structure G_{out} for the target organism. The procedure can be repeated for both organisms using the output DAG of one organism as an input for the other organism.

Kato et al. [45] considered multiple assays where learning via the sharing of local knowledge occurs, reflected by learning weights in the formula below:

$$\xi(\{w_t\}) = \sum_{t=1}^{TN}\sum_{i=1}^{n_t} l(y_i^t, w_t^T x_i^t) + \lambda_1\sum_{t=1}^{TN}\|w_t\|_2^2$$
$$+ \lambda_2\sum_{t=1}^{TN}\sum_{v \in V_t}(\|w_v\|^2 + C\|w_c - w_t\|) \qquad (13)$$

where V_t represents the local model of a target node t. Intuitively, this formulation employs transfer learning of the target task with the help of its neighbors' tasks.

Qi et al. [46] proposed a semi-supervised multi-task framework to predict protein-protein interactions from partially labeled reference sets. The basic idea is to perform multitask learning on a supervised classification task and a semi-supervised auxiliary task via a regularization term. This is equivalent to learning two tasks jointly while optimizing the following loss function:

$$\xi(\{w\}) = \sum_{i=1}^{n_{labeled}} l(y_i, w^T x_i) + Loss(Auxiliary\ Task) \qquad (14)$$

Protein-protein interaction prediction is an important aspect of systems biology. Besides the work of Qi et al. [46], Xu et al. [47] also explored how to use multitask learning in this area via a technique known as Collective Matrix Factorization (CMF) [48]. These methods use similarities of the proteins in two interaction networks as the corresponding shared knowledge. They showed that when the source matrix is sufficiently dense and similar to the target PPI network, transfer learning is effective for predicting protein-protein interactions in a sparse network. Consider a similarity matrix $S_{m \times n}$ introduced as the correspondence between networks G and P. The rows and columns of $S_{m \times n}$ correspond to proteins in networks G and P, respectively, and the element S_{ij} of $S_{m \times n}$ represents the similarity between node i in network G and node j in network P. The collective matrix

factorization method reconstructs matrices $X^t \approx f_1(ZV^T)$ and $X^\alpha \approx f_2(UV^T)$ together by sharing the common factor V. The objective of collective matrix factorization then is to minimize the regularized loss:

$$\zeta(X^t, X^\alpha, U, V, Z) = D(X^t \| ZV^T) + \lambda^s D(X^\alpha \| UV^T) + \lambda^U \|U\|_F^2$$
$$+ \lambda^V \|V\|_F^2 + \lambda^Z \|Z\|_F^2 \qquad (15)$$

where

$$X^t = \begin{bmatrix} L_{m \times m} & 0 \\ 0 & L_{n \times n} \end{bmatrix}, \quad X^\alpha = \begin{bmatrix} 0 & S_{m \times n} \\ S^t_{m \times n} & 0 \end{bmatrix}$$

A sparse multitask regression approach was presented in [49], where a co-clustering algorithm is applied to gene expression data with phenotypic signatures. This algorithm can uncover the dependency between genes and phenotypes. A multitask learning framework with L_1 norm regularization is given as follows

$$\xi(\{T, P_d\}) = \sum_{d=0}^{D} \|X_0 T P_d - Y_d\|_F^2 + \lambda \|T\|_1 \qquad (16)$$

$$s.t. \ \|P_d\|_F = 1, \text{ for } d = 1,2,..., D$$

where $T_d = T \cdot P_d$, which represents phenotype responses under different experimental conditions being projected on the same low-dimensional space T. In this equation, the first term enforces a fit between the gene expression and the phenotypic signature under each condition, while the second term enforces sparsity on T.

Bickel et al. [50] studied the problem of predicting HIV therapy outcomes of different drug combinations based on observed genetic properties of the patients, where each task corresponds to a particular drug combination. They proposed to jointly train models of different drug combinations by pooling data together from all tasks and then use re-sampling weights to adapt the data for each particular task. The goal is to learn a hypothesis $f_t : x \rightarrow y$ for each task t by minimizing a loss function with respect to $p(x,y|t)$, where x describes the genotype of the virus that a patient carries, together with the patient's treatment history, y is a class label indicating whether the therapy is successful. Simply pooling the available data for all tasks will generate a training sample $D = \langle (x_1,y_1,t_1),...,(x_{TN},y_{TN},t_{TN}) \rangle$. The approach creates a task-specific re-sampling weight $r_t(x,y)$ for each element in the pool of examples.

V. BIOMEDICAL APPLICATIONS

Text mining provides rich ground for transfer learning and multi-task learning applications. In this area, several approaches to general text mining with transfer learning have been surveyed in [3].

In the biomedical domain, one important area is semantic role labeling (SRL) systems, these label the roles of genes, proteins and biological entities discussed in textual form. These texts are often labeled based on manually annotated training instances, which are rare and expensive to prepare. To solve this problem, Dahlmeier and Ng [51] formulated SRL in the biomedical domain as a transfer learning problem, to leverage existing SRL resources from the newswire domain. They

employed three domain transfer learning methods: instance weighting, the augment method and instance pruning. Instance weighting [6, 52] is aimed at correcting the probability estimate for the target domain by weighting instances that occur in the auxiliary data sets. Domain adaptation methods [53] map features from the auxiliary and target domains to a common feature space where knowledge transfer is possible.

In addition to biomedical text mining applications of transfer learning, Bi et al. [54] formulated the detection of different types of clinically-related abnormal structures in medical images using multitask learning. Their method captured task-dependence by sharing common feature representations, this was shown to be effective in eliminating irrelevant features and identifying discriminative features. Given TN tasks, for each task t, the sample set is composed of input features X^t and label vector y^t. Traditional single-task learning aims to minimize the objective function: $L(a_t, X^t, y^t) + \lambda P(\alpha_t)$ for each task individually, where the former term is a loss function and the latter term regularizes the complexity of the model. In [54], a family of jointly learning algorithms can be derived by rewriting $\alpha_t = C\beta_t$, where β_t is task-specific while C is a diagonal matrix with a diagonal vector equal to $c \geq 0$. The objective function can be rewritten as:

$$\xi(\{\beta_t\}) = \sum_{t=1}^{TN} (l(C\beta_t, X^t, y^t)) + P_1(\beta_t)) \qquad (17)$$

$$s.t. P_2(c) \leq \gamma$$

where P_1 and P_2 are regularization operators, and C is a diagonal matrix with a diagonal vector equal to $c \geq 0$. In other words, c is an indicator vector showing if a feature is used in the model, which is then used as a common feature representation across all tasks.

In the field of sensor-based ubiquitous healthcare, and particularly in motion-sensor-based activity recognition, collecting user labeled samples needs huge manual efforts and may involve privacy issues. Therefore, transfer learning becomes an attractive approach to solving the data sparsity problem. Here we give one example of such a work, which transferred the activity models for a new user by transfer learning [55]. In earlier work, transfer learning was employed to transfer the activity models learned for a user to another user. Van Kasteren et al. [56] described a simple method for transferring the transitional probabilities of two Markov models for two different spaces. In this work, the aim is to recognize activities in a new space without the expensive data annotation and learning processes. Kasteren et al. [56] first mapped data between two sensor networks via a mapping function. Then their algorithm learned parameters using the labeled data in the source space and unlabeled data in the target space. In this work, the mapping is generated manually and the structure of the HMMs is pre-defined. However, in practice, the mapping and model structure should be learned.

To address this problem, Rashidi and Cook [57] proposed an advanced transfer learning approach to transfer activity knowledge learned in a home to another home, in what they call "Home to Home Transfer Learning" (HHTL). The main components of HHTL are illustrated in the Fig. 4. As the figure shows, HHTL extracts and compresses activity models from the sensor
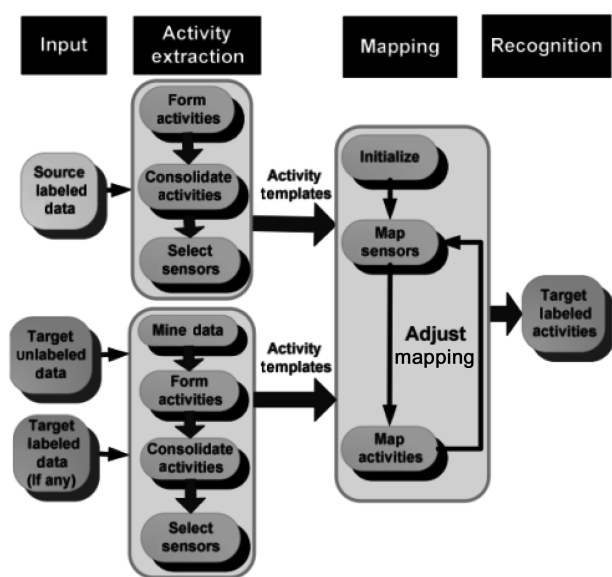
Fig. 4. Main components of Home to Home Transfer Learning (HHTL) for transferring activities from a source space to a target space (Rashidi and Cook [57]).

data based on label information in a source domain where data is assumed to have been collected and labeled. This data can then be used to build activity models that transfer to the target domain. After extracting activities and their corresponding sensors, source activity models are mapped to target activity models using a semi-EM framework in an iterative manner. Initially, sensor mapping probabilities are adjusted based on activity mapping probabilities. Next, the activity mapping probabilities are adjusted based on the updated sensor mapping probabilities. A target activity's label is determined to be the same as the source activity's label when the mapping probability is maximized.

Subsequently, Rashidi and Cook [58] modified the HHTL algorithm to transfer the knowledge of learned activities from multiple source physical spaces. E.g., knowledge transfer is done from two homes A and B, to a target physical space, e.g. home C. The modified learning model (MHTL) has the same first three steps illustrated in Fig. 4, but in the final step, MHTL adopts a weighted majority voting scheme to assign the final labels to activities in the target domain.

VI. DATA SETS FOR TRANSFER LEARNING

Several datasets have been released for further research on transfer learning in biological domains. We list several datasets available here.

Gene Expression Data: Yeast gene expression data and human gene expression data are available from ([59], http://genome-www.stanford.edu/cellcycle/) and ([60], http://www.stanfore.edu/Human-CellCycle/Hela/), respectively. These two sets of gene-expression data can be used to reconstruct gene networks across different organisms.

Protein-ligand Binding Data: One group of available data is composed of enzyme-ligand interaction data, G-protein-coupled receptors (GPCR)-ligand interaction data and ion channel-ligand interaction data (http://bioinformatics.oxfordjournals.org/content/24/19/2149/suppl/DC1). Table 1 lists the statistics of the data. Another group of released ligand interaction data contains four subsets for enzyme, ion channel, GPCR and nuclear receptors [61] (http://web.kuicr.kyoto-u.ac.jp/supp/yoshi/drug-target/), respectively. The ligand structure similarity matrices and protein sequence similarity matrices are also included in these datasets. Interested readers can build models to predict the protein-ligand binding via a multi-task framework using these datasets or predict the protein-ligand binding via transferring knowledge from auxiliary data such as the ligand structure similarity matrices and protein sequence similarity matrices.

HIV-1 and Human Interaction Data: Qi et al. [46] preprocessed a HIV-1 and protein interaction dataset (http://www.cs.cmu.edu/qyj/HIVsemi/), and exploited a semi-supervised multi-task method to predict HIV-1 and protein interactions.

Protein Interaction Data: Multi-task learning methods for reconstructing biological networks can be tested on metabolic network and protein interaction network data, not just on gene interaction related data. Gold standard metabolic network data, the corresponding gene expression data and cross-specie sequence similarities for *C. elegans*, *H. pylor,* and *S. cerevisi-aewere* released in [62] (http://cbio.ensmp.fr/ yyamanishi/Link-Propagation/).

Splice Site Data: Acceptor splicing site data for *C. elegans, C. remanei, P. pacificus, D. melanogaster,* and *A. thaliana* are available (http://www.fml.tuebingen.mpg.de/raetsch/suppl/genom-edomainadaptation).

VII. SUMMARY

Transfer learning has been employed in various bioinformatics applications and has attracted more attention over recent years. Below we briefly summarize some bioinformatics applications using transfer learning techniques (Table 2).

VIII. CONCLUSION AND FUTURE WORK

In this article, we reviewed research work in transfer learning and multitask learning in the area of bioinformatics and bio-

Table 1. Statistics of datasets

	No. of proteins	No. of compounds	No. of training data
Enzymes	2,436	675	524
G-protein-coupled receptors	798	100	219
Ion channel	2,330	114	462

Table 2. Applications in bioinformatics

Biological sequence analysis	Splicing site recognition	Widmer et al. [34], Schweikert et al. [31], etc.
	MHC binding prediction	Jacob and Vert [34], Jacob et al. [35], etc.
	siRNA efficacy prediction	Liu et al. [39], etc.
	Protein subcellular localization	Xu et al. [36], Mei et al. [63], etc.
Genetic data analysis	Gene expression data analysis	Chen and Huang [40], Xu et al. [42], etc.
	Genetic analysis	Puniyani et al. [41], etc.
Systems biology	Gene network reconstruction	Tamada et al. [43], Nassar et al. [44], Kato et al. [45], etc.
	Protein-protein interaction network reconstruction	Qi et al. [46], Xu et al. [47], etc.
	Drug-protein interaction prediction	Zhang et al. [49], Bickel et al. [50], Jacob and Vert [64], etc.
Biomedical applications	Biomedical text mining	Dahlmeier and Ng [51], etc.
	Bioimage mining	Bi et al. [54], etc.
	Ubiquitous Healthcare	Zhao et al. [65], Van Kasteren et al. [56], Rashidi and Cook [55, 57, 58], etc.

MHC: major histocompatibility complex, siRNAs: small interfering RNAs.

medical applications. These works included regularization based methods, common feature representation based methods, distribution matching methods and their variants. Transfer learning has attracted more and more attention in scientific communities, and has matured over the last few years. Researchers have begun to investigate and explore many other applications in bioinformatics using transfer learning. As we have seen, most current work on transfer learning is focused on sequence classification, gene expression data analysis, biological network reconstruction, biomedical text, image mining, and sensor-based ubiquitous healthcare. This research aims to alleviate the data sparsity problem in biological domains. An important example of real-life transfer learning in bioinformatics is the GRAIL system (http://www.broadinstitute.org/mpg/grail/), where transfer learning is conducted across feature spaces. In the future, more transfer learning applications in bioinformatics are expected, especially when new data is generated that shares some biological properties with existing data, where we have already gained significant insights. Such insights can then be propagated to new application domains through transfer learning.

ACKNOWLEDGMENT

REFERENCES

1. P. Larranaga, B. Calvo, R. Santana, C. Bielza, J. Galdiano, I. Inza, J. A. Lozano, R. Armananzas, G. Santafe, A. Perez, and V. Robles, "Machine learning in bioinformatics," *Brief Bioinformatics*, vol. 7, no. 1, pp. 86-112, Mar. 2006.

2. R. A. Caruana, "Multitask learning: a knowledge-based source of inductive bias," *Proceedings of theTenth International Conference on Machine Learning*, Amherst, MA, 1993, pp. 41-48.

3. S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345-1359, 2010.

4. J. Blitzer, R. Mcdonald, and F. Pereira, "Domain adaptation with structural correspondence learning," *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Sydney, Australia, 2006, pp. 120-128.

5. C. B. Do and A. Y. Ng, "Transfer learning for text classification," *Proceedings of the 19th Annual Conference on Neural Information Processing Systems*, Vancouver, Canada, 2005.

6. W. Dai, Q. Yang, G. R. Xue, and Y. Yu, "Boosting for transfer learning," *Proceedings of the 24th International Conference on Machine Learning*, Corvalis, OR, 2007, pp. 193-200.

7. P. C. Woodland, "Speaker adaptation for continuous density HMMS: a review," *ISCA Tutorial and Research Workshop on Adaptation Methods for Speech Recognition*, Sophia Antipolis, France, 2001.

8. L. Xiao and J. Bilmes, "Regularized adaptation of discriminative classifiers," *IEEE International Conference on Acoustics, Speech and Signal Processing*, Toulouse, France, 2006, pp. I237-I240.

9. R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng, "Self-taught learning: transfer learning from unlabeled data," *Proceedings of the 24th International Conference on Machine Learning*, Corvalis, OR, 2007, pp. 759-766.

10. J. Yang, R. Yan, and A. G. Hauptmann, "Cross-domain video concept detection using adaptive SVMs," *Proceedings of the 15th ACM International Conference on Multimedia*, Augsburg, Germany, 2007, pp. 188-197.

11. V. W. Zheng, D. H. Hu, and Q. Yang, "Cross-domain activity recognition," *Proceedings of the 11th ACM International Conference on Ubiquitous Computing*, Orlando, FL, 2009, pp. 61-70.

12. H. Y. Wang, V. W. Zheng, J. Zhao, and Q. Yang, "Indoor localization in multi-floor environments with reduced effort," *Proceedings of the 8th IEEE International Conference on Pervasive Computing and Communications*, Mannheim, Germany, 2010, pp. 244-252.

13. P. Donnes and A. Elofsson, "Prediction of MHC class I binding

peptides, using SVMHC," *BMC Bioinformatics*, vol. 3 pp. 25, Sep. 2002.

14. K. Aas, Microarray Data Mining: A Survey, Oslo, Norway: Norwegian Computing Center, 2001.

15. E. P. Xing, M. I. Jordan, and R. M. Karp, "Feature selection for high-dimensional genomic microarray data," *Proceedings of the 8th International Conference on Machine Learning*, Williamstown, MA, 2001, pp. 601-608.

16. W. H. Yang, D. Q. Dai, and H. Yan, "Finding correlated biclusters from gene expression data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, no. 4, pp. 568-584, 2011.

17. C. Yang, Z. He, X. Wan, Q. Yang, H. Xue, and W. Yu, "SNPHarvester: a filtering-based approach for detecting epistatic interactions in genome-wide association studies," *Bioinformatics*, vol. 25, no. 4, pp. 504-511, Feb. 2009.

18. X. Wan, C. Yang, Q. Yang, H. Xue, N. L. Tang, and W. Yu, "MegaSNPHunter: a learning approach to detect disease predisposition SNPs and high level interactions in genome wide association study," *BMC Bioinformatics*, vol. 10 pp. 13, 2009.

19. M. Krallinger and A. Valencia, "Text-mining and information-retrieval services for molecular biology," *Genome Biology*, vol. 6, no. 7, pp. 224, 2005.

20. C. Orwat, A. Graefe, and T. Faulwasser, "Towards pervasive computing in health care: a literature review," *BMC Medical Informatics and Decision Making*, vol. 8 pp. 26, 2008.

21. L. Bao and S. S. Intille, "Activity recognition from user-annotated acceleration data," *Pervasive Computing: Second International Conference, PERVASIVE 2004, Linz/Vienna, Austria, April 21-23, 2004. Proceedings. Lectures Note in Computer Science Vol. 3001*, A. Ferscha and F. Mattern, Eds., Heidelberg: Springer Berlin, 2004, pp. 1-17.

22. N. Ravi, N. D, P. Mysore, and M. L. Littman, "Activity recognition from accelerometer data," *Proceedings of the 17th Conference on Innovative Applications of Artificial Intelligence*, Pittsburgh, PA, 2005, pp. 1541-1546.

23. J. Yin, X. Chai, and Q. Yang, "High-level goal recognition in a wireless LAN," *Proceedings of the 19th National Conference on Artificial Intelligence; Sixteenth Innovative Applications of Artificial Intelligence Conference*, San Jose, CA, 2004, pp. 578-583.

24. S. J. Cho, J. K. Oh, W. C. Bang, W. Chang, E. Choi, Y. Jing, J. Cho, and D. Y. Kim, "Magic wand: a hand-drawn gesture input device in 3-D space with inertial sensors," *Ninth International Workshop on Frontiers in Handwriting Recognition*, Tokyo, Japan, 2004, pp. 106-111.

25. M. Chen, B. Huang, and Y. Xu, "Human abnormal gait modeling via hidden Markov model," *Proceedings of the 2007 International Conference on Information Acquisition*, Jeju, Korea, 2007, pp. 517-522.

26. S. D. Choi, A. S. Lee, and S. Y. Lee, "On-line handwritten character recognition with 3D accelerometer," *Proceedings of the IEEE International Conference on Information Acquisition*, Weihai, China, 2006, pp. 845-850.

27. L. Nachman, A. Baxi, S. Bhattacharya, V. Darera, P. Deshpande, N. Kodalapura, V. Mageshkumar, S. Rath, J. Shahabdeen, and R. Acharya, "Jog falls: a pervasive healthcare platform for diabetes management," *Pervasive Computing. Lecture Notes in Computer Science Vol. 6030*, P. Floreen, A. Kruger, and M. Spasojevic, Eds., Heidelberg: Springer Berlin, pp. 94-111, 2010.

28. S. Consolvo, K. Everitt, I. Smith, and J. A. Landay, "Design

requirements for technologies that encourage physical activity," *Conference on Human Factors in Computing Systems*, Montreal, Canada, 2006, pp. 457-466.

29. T. Evgeniou and M. Pontil, "Regularized multi-task learning," *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Seattle, WA, 2004, pp. 109-117.

30. C. Widmer, J. Leiva, Y. Altun, and G. Ratsch, "Leveraging sequence classification by taxonomy-based multitask learning," *Research in Computational Molecular Biology: 14th Annual International Conference, RECOMB 2010, Lisbon, Portugal, April 25-28, 2010. Proceedings. Lecture Notes in Computer Science Vol. 6044*, B. Berger, Ed., Heidelberg: Springer Berlin, pp. 522-534, 2010.

31. G. Schweikert, C. Widmer, B. Scholkopf, and G. Ratsch, "An empirical analysis of domain adaptation algorithms for genomic sequence," *Proceedings of the 22nd Annual Conference on Neural Information Processing Systems*, Vancouver, Canada, 2008.

32. L. Jacob and J. P. Vert, "Efficient peptide-MHC-I binding prediction for alleles with few known binders," *Bioinformatics*, vol. 24, no. 3, pp. 358-366, Feb. 2007.

33. L. Jacob, F. Bach, and J. P. Vert, "Clustered multi-task learning: a convex formulation," *Proceedings of the 23th Annual Conference on Neural Information Processing Systems*, Vancouver, Canada, 2009.

34. C. Widmer, N. Toussaint, Y. Altun, O. Kohlbacher, and G. Ratsch, "Novel machine learning methods for MHC class I binding prediction," *Pattern Recognition in Bioinformatics: 5th IAPR International Conference, PRIB 2010, Nijmegen, The Netherlands, September 22-24, 2010. Proceedings. Lecture Notes in Computer Science Vol. 6282*, T. Dijkstra, E. Tsivtsivadze, E. Marchiori, and T. Heskes, Eds., Heidelberg: Springer Berlin, pp. 98-109, 2010.

35. C. Widmer, N. C. Toussaint, Y. Altun, and G. Ratsch, "Inferring latent task structure for Multitask Learning by Multiple Kernel Learning," *BMC Bioinformatics*, vol. 11 Suppl 8 pp. S5, 2010.

36. Q. Xu, S. J. Pan, H. H. Xue, and Q. Yang, "Multitask learning for protein subcellular location prediction," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 8, no. 3, pp. 748-759, 2011.

37. A. Argyriou, T. Evgeniou, and M. Pontil, "Multi-task feature learning," *Advances in Neural Information Processing Systems Vol. 19: Proceedings of the 2006 Neural Information Processing System Conference*, B. Scholkopf, J. Platt, and T. Hofmann, Eds., Cambridge, MA: MIT Press, 2007, pp. 41-48.

38. A. Argyriou, T. Evgeniou, and M. Pontil, "Convex multi-task feature learning," *Machine Learning*, vol. 73, no. 3, pp. 243-272, 2008.

39. Q. Liu, Q. Xu, V. W. Zheng, H. Xue, Z. Cao, and Q. Yang, "Multi-task learning for cross-platform siRNA efficacy prediction: an in-silico study," *BMC Bioinformatics*, vol. 11 pp. 181, 2010.

40. A. Chen and Z. W. Huang, "A new multi-task learning technique to predict classification of leukemia and prostate cancer," *Medical Biometrics. Lecture Notes in Computer Science Vol. 6165*, D. Zhang and M. Sonka, Eds., Heidelberg: Springer Berlin, pp. 11-20, 2010.

41. K. Puniyani, S. Kim, and E. P. Xing, "Multi-population GWA mapping via multi-task regularized regression," *Bioinformatics*,

vol. 26, no. 12, pp. i208-i216, Jun. 2010.

42. Q. Xu, H. Xue, and Q. Yang, "Multi-platform gene expression mining and marker gene analysis," *International Journal of Data Mining and Bioinformatics* 2010 in press.

43. Y. Tamada, H. Bannai, S. Imoto, T. Katayama, M. Kanehisa, and S. Miyano, "Utilizing evolutionary information and gene expression data for estimating gene networks with Bayesian network models," *Journal of Bioinformatics and Computational Biology*, vol. 3, no. 6, pp. 1295-1313, 2005.

44. M. Nassar, R. Abdallah, H. A. Zeineddine, E. Yaacoub, and Z. Dawy, "A new multitask learning method for multiorganism gene network estimation," *IEEE International Symposium on Information Theory*, Toronto, Canada, 2008, pp. 2287-2291.

45. T. Kato, K. Okada, H. Kashima, and M. Sugiyama, "A transfer learning approach and selective integration of multiple types of assays for biological network inference," *International Journal of Knowledge Discovery in Bioinformatics*, vol. 1, no. 1, pp. 66-80, Jan.-Mar. 2010.

46. Y. Qi, O. Tastan, J. G. Carbonell, J. Klein-Seetharaman, and J. Weston, "Semi-supervised multi-task learning for predicting interactions between HIV-1 and human proteins," *Bioinformatics*, vol. 26, no. 18, pp. i645-i652, 2010.

47. Q. Xu, E. W. Xiang, and Q. Yang, "Protein-protein interaction prediction via collective matrix factorization," *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine*, Hong Kong, 2010, pp. 62-67.

48. A. P. Singh and G. J. Gordon, "Relational learning via collective matrix factorization," *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Las Vegas, NV, 2008, pp. 650-658.

49. K. Zhang, J. W. Gray, and B. Parvin, "Sparse multitask regression for identifying common mechanism of response to therapeutic targets," *Bioinformatics*, vol. 26, no. 12, pp. i97-i105, Jun. 2010.

50. S. Bickel, J. Bogojeska, T. Lengauer, and T. Scheffer, "Multi-task learning for HIV therapy screening," *Proceedings of the 25th International Conference on Machine Learning*, Helsinki, Finland, 2008, pp. 56-63.

51. D. Dahlmeier and H. T. Ng, "Domain adaptation for semantic role labeling in the biomedical domain," *Bioinformatics*, vol. 26, no. 8, pp. 1098-1104, 2010.

52. J. Jiang and C. Zhai, "Instance weighting for domain adaptation in NLP," *Proceedings of the 45th Annual Meeting of the Associationfor Computational Linguistics*, Prague, Czech Republic, 2007, pp. 264-271.

53. H. Daume III, "Frustratingly easy domain adaptation," *Proceedings of the 45th Annual Meeting of the Associationfor Computational Linguistics*, Prague, Czech Republic, 2007, pp. 256-263.

54. J. Bi, T. Xiong, S. Yu, M. Dundar, and R. Rao, "An improved multi-task learning approach with applications in medical diagno-

sis," *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2008, Antwerp, Belgium, September 15-19, 2008, Proceedings, Part I. Lecture Notes in Computer Science Vol. 5211*, W. Daelemans, B. Goethals, and K. Morik, Eds., Heidelberg: Springer Berlin, pp. 117-132, 2008.

55. P. Rashidi and D. J. Cook, "Transferring learned activities in smart environments," *Proceedings of the 5th International Conference on Intelligent Environments*, Barcelona, Spain, 2009, pp. 185-192.

56. T. L. M. Van Kasteren, G. Englebienne, and B. J. A. Krose, "Recognizing activities in multiple contexts using transfer learning," *AAAI Fall Symposium Technical Report Vol. FS-08-02*, Arlington, VA, 2008, pp. 142-149.

57. P. Rashidi and D. J. Cook, "Home to home transfer learning," *Proceedings of the AAAI Plan, Activity, and Intent Recognition Workshop*, Atlanta, GA, 2010.

58. P. Rashidi and D. J. Cook, "Multi home transfer learning for resident activity discovery and recognition," *Proceedings of the 4th International Workshop on Knowledge Discovery from Sensor Data*, Washington, DC, 2010.

59. P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher, "Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization," *Molecular Biology of the Cell*, vol. 9, no. 12, pp. 3273-3297, Dec. 1998.

60. M. L. Whitfield, G. Sherlock, A. J. Saldanha, J. I. Murray, C. A. Ball, K. E. Alexander, J. C. Matese, C. M. Perou, M. M. Hurt, P. O. Brown, and D. Botstein, "Identification of genes periodically expressed in the human cell cycle and their expression in tumors," *Molecular Biology of the Cell*, vol. 13, no. 6, pp. 1977-2000, Jun. 2002.

61. Y. Yamanishi, M. Araki, A. Gutteridge, W. Honda and M. Kanehisa, "Prediction of drug-target interaction networks from the integration of chemical and genomic spaces", *Bioinformatics*, vol. 24, no. 13, pp. i232-i240, 2008.

62. H. Kashima, Y. Yamanishi, T. Kato, M. Sugiyama, and K. Tsuda, "Simultaneous inference of biological networks of multiple species from genome-wide data and evolutionary information: a semi-supervised approach," *Bioinformatics*, vol. 25, no. 22, pp. 2962-2968, Nov. 2009.

63. S. Mei, W. Fei, and S. Zhou, "Gene ontology based transfer learning for protein subcellular localization," *BMC Bioinformatics*, vol. 12 pp. 44, 2011.

64. L. Jacob and J. P. Vert, "Protein-ligand interaction prediction: an improved chemogenomics approach," *Bioinformatics*, vol. 24, no. 19, pp. 2149-2156, Oct. 2008.

65. Z. Zhao, Y. Chen, J. Liu, and M. Liu, "Cross-mobile ELM based activity recognition," *International Journal of Engineering and Industries*, vol. 1, no. 1, pp. 30-38, Dec. 2010.
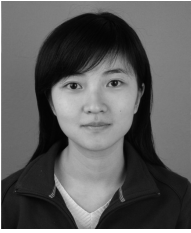
Qian Xu

Qian Xu is currently a postdoc fellow in the Department of Computer Science and Engineering, Hong Kong University of Science and Technology. She got her Ph.D degree at the Bioengineering Program of Hong Kong University of Science and Technology.

Qiang Yang

Qiang Yang is a professor in the Department of Computer Science and Engineering, Hong Kong University of Science and Technology and an IEEE Fellow. His research interests are data mining and artificial intelligence (including automated planning, machine learning and activity recognition). He received his PhD from Computer Science Department of the University of Maryland, College Park in 1989. He was an assistant/associate professor at the University of Waterloo between 1989 and 1995, and a professor and NSERC Industrial Research Chair at Simon Fraser University in Canada from 1995 to 2001. His research teams won the 2004 and 2005 ACM KDDCUP international competitions on data mining. He is an invited speaker at IJCAI 2009, ACL 2009, ACML 2009 and ADMA 2008, etc. He was elected as a vice chair of ACM SIGART in July 2010. He is the founding Editor in Chief of the ACM Transactions on Intelligent Systems and Technology (ACM TIST), and is on the editorial board of IEEE Intelligent Systems and several other international journals. He was a PC co-chair or general co-chair for a number of international conferences, including ACM KDD 2010, ACM IUI 2010, ACML 2010, PRICAI 2006, PAKDD 2007, ADMA 2009, ICCBR 2001, etc. He serves on the IJCAI 2011 advisory committee.