Overview of Integrative Analysis Methods for Heterogeneous Data

(Invited Paper)

Jaya Thomas
Department of Computer Science,
[1]State University of New York, Korea,
Incheon 406-840, Korea
[2]Stony Brook University, Stony Brook,
NY 11794, USA;
Email:jaya.thomas@sunykorea.ac.kr

Lee Sael
Department of Computer Science,
[1]State University of New York, Korea,
Incheon 406-840, Korea
[2]Stony Brook University, Stony Brook,
NY 11794, USA;
Email: sael@sunykorea.ac.kr

***Abstract*—In the big data era, data are not only generated in massive quantity but also in diversity. The heterogeneous characteristics of the diverse data sources on a subject provide complimentary information. However, they pose challenges in data analysis process. Then, what are the existing methods for utilizing theses heterogeneous data to improve data analysis and how can we choose amongst these methods? We categorize integrative methods for heterogeneous data analysis to Bayesian network based methods and multiple kernel based methods and describe them in detail with examples of successful applications in the bioinformatics field.**

I. Introduction

In the big data era, we are not only faced with the massiveness of data but also with heterogeneity of the data that often provide complementary views about a subject. We can expect that having more information about a subject, we have better chance of analyzing it with higher accuracy. However, analysis of these data are often challenging due to inconsistencies in the data that results from diversity in data extraction environment and perspectives. This emerges the need for integrative analysis methods that are able to analyze heterogeneous data, i.e., methods that are able to combines data with diverse background distributions, relations, dimensions, and formats to enhance the statistical significance and obtain more refined information.

The heterogeneity may result due to various reasons. It may be due to difference in data extraction environment and what perspective of the subject is being studied. Difference in the data extraction environment includes difference in the data sources, and the extraction methods. Data source depends on the research question being addressed. In case of biological experiments, a source can be a model organism or a cell line used to carry out the experiment. There is also diversity in the data extraction methods and they often result in different background distributions. The perspective of the study being performed on the subject also contributes to the heterogeneity in various aspects including what is being measured, type of experiment performed, and resolution of the data.

Considering data formats, there are heterogeneity in measurement scales, dimensions, and types. There is wide spectrum of data formats used to represent data ranging from high-resolution images, structured, multi-dimensional data to networks, vector, etc. Different formats require different data types (numeric, character, textual, and etc.) to store the data. They are also associated with measurement scale of data: nominal, ordinal, interval and ratio.

The heterogeneous data integration methods are being explored in numerous fields of study. Some of the successful application includes integrative methods used for gene prioritization [1], bacteria classification and gene function prediction [2], siRNA efficacy prediction [3], signal processing applications [4], visual object recognition[5], protein function prediction [6], and inference of patient-specific pathway activities [7].

There are many applications that attempt to integrate heterogeneous many of which are heuristic approaches that depend heavily on the specific problem being targeted. Although some methods are difficult to be generalized, we find that there are two major class of methods that are explored for heterogeneous data analysis: Bayesian network based methods and multiple kernel based methods.

II. Network Based Methods

A. Bayesian Networks

Bayesian networks (BN) are one of the parametric learning methods that the data are assumed to be drawn from a probability distribution of specific parameter values. More formally, BN is a directed acyclic graph *G* that represents the joint probability distribution over the set of random variables X_1,X_2,. . .,X_n where the nodes represents the variable and edges represents the conditional dependencies between the variables. The graph structure of BN is capable of representing a joint probability distribution of a domain as

$$P_r(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P_r(X_i|Pa_i) \qquad (1)$$

where Pa_i denotes the parents of X_i. Thus, the joint probability distribution can be factored into smaller local probability distribution each involving a node and its parents.

978-1-4799-7303-3/15/$31.00 ©2015 IEEE

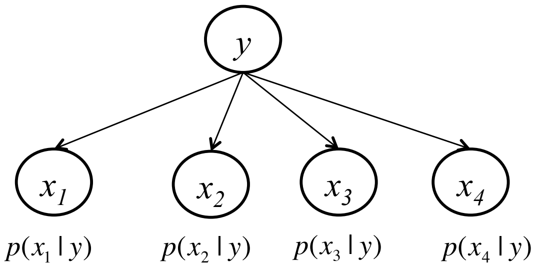

Fig. 1. Example of naive bayes model for integrating data sources.

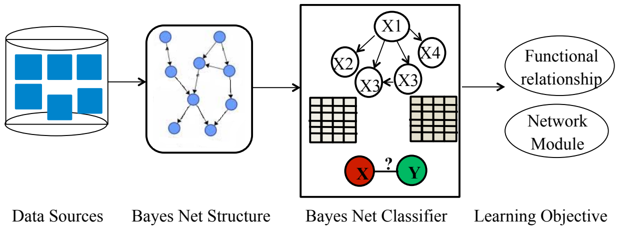

Fig. 2. Example Bayesian network framework for heterogeneous data integration.

BNs are often used to represent complex relationships among variables for several reasons. First of all, they can handle uncertainty in knowledge explicitly. In addition, BN's graphical representation provides an intuitive method for integrating existing knowledge about the variables. The graphical structure of BN enforces certain dependency constraints as shown in Fig. 1.

BNs are also ideal for integrative analysis of heterogeneous data as they not only provide the means to model relations between variables, they can also be extended to model relations between heterogeneous data of each variable. A simplest application of BN of integrating several types of data or observations would be to assume independence between observations (xs in Fig. 1) and use naive Bayes model to determine the set of data infers a hidden factor (y in Fig 1) [6]. More complex application of BN will model the relationship between observations [7].

In modeling BN for heterogeneous data integration, two components are considered:

- network structure in a form of directed acyclic graph (DAG) that represents relationship between different observations (heterogeneous data) of variables that describe the subjet and relationship between the variables themselves, and
- set of the local probability distributions one for each variable and its observed characters, conditioned on each value combination of the parents.

Applications of the BN on heterogeneous data integration follow the same steps of the regular BN analysis. First step is construction of BN structure. There are three approaches to obtain the structure of BN: 1) manual construction based on expert knowledge, 2) structure learning using the massive data, and 3) mixture of the two approaches. After construction of BN structure, parameters associated with the conditional probabilities can be learned by maximum likelihood approaches. When a new data is processed, various inference algorithms including sampling methods and belief propagation approaches can make inference about a variable in the BN.

B. Applications of Bayesian Networks on Heterogeneous Data Integration

Bayesian network (BN) provides a flexible framework for integration heterogeneous data and researches of various flavors have been performed. We look at some of examples of how BN can be applied in the heterogeneous data integration by looking at example applications.

1) Example that learn both structure and parameter of BN: Gevaert *et al.* [8] presented a BN framework to prognosis breast cancer by integrating clinical and microarray data with BN. They proposed and evaluated three different integration methods, namely full integration, partial integration and decision integrations based on the time point of the integration. In the full integration, integration occurs in the initial stage where the two data types (clinical and microarray data) are merged and processed as one dataset. In the partial integration, BN structures for each data type are learned first and then integrated to form one structure. The parameter learning is done on the integrated structure. In the decision integration, the predictions of learned models for each data types are integrated after individual data. Out of the three methods used to test the prediction accuracy, partial and decision integration was shown to performs better on heterogeneous data as compared to full integration method and individual data evaluations.

In their work, learning process of BN model was developed in two steps: structural learning and parameter learning [8]. The structure of the BN was learned using K2 [9], a greedy search algorithm, based on Bayesian Dirichlet (BD) scoring metric [10]. The BD scoring metric is shown in the following equation:

$$p(S|D) \propto p(S) \prod_{i=1}^{n} \prod_{j=1}^{q_i} \left[\frac{\Gamma(N^{'}_{ij})}{\Gamma(N^{'}_{ij} + N_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(N^{'}_{ijk} + N_{ijk})}{\Gamma(N^{'}_{ijk})} \right] \qquad (2)$$

where S denotes the current structure, N_{ijk} are the number of cases in the dataset D (containing n variables) having variable i in the state k with the j^{th} instantiation of its parents. The r_i is the number of state of variable i and being q_i the number of instantiation of the parent of variable i. $N^{'}_{ij}$ refers to prior knowledge of parameter and N_{ij} is computed as given by Eq. 3 and in cases no prior knowledge is available, it is computed using Eq. 4. Γ (.) denotes the gamma distribution and represents prior structure probability. The K2 search strategy use eq. 2 to score the structure.

$$N_{ij} = \sum_{k=1}^{r_i} N_{ijk}.N^{'}_{ijk}. \qquad (3)$$

$$N^{'}_{ijk} = N/(r_i q_i). \qquad (4)$$

After generation of BNs for each data types, parameters of conditional probability table (CPT) based on uniform Dirichlet prior is learned. The uniform Dirichlet prior is formulated as follows:

$$p(\Theta_{ij}|S) = Dir(\Theta_{ij}|N^{'}_{ij1}, \cdots, N^{'}_{ijk}, \cdots, N^{'}_{ijr_i}, S) \tag{5}$$

where Θ_{ij} corresponds to parameter set, where it contains the probability for every value of the variable X$_i$ given the current instantiation of the parent. The parameter for Dirichlet are chosen un-informatively and are updated with the data that results in Dirichlet posterior over the parameter set denoted by Eq. 6. The CPT for each variable is computed by Maximum A Posteriori (MAP) parameterization of the Dirichlet distribution.

$$\begin{aligned} p(\Theta_{ij}|D, S) = Dir(\Theta_{ij}|N^{'}_{ij1} + N_{ij1}, \cdots, N^{'}_{ijk} + N_{ijk}, \\ \cdots, N^{'}_{ijr_i} + N_{ijr_i}, S). \end{aligned} \tag{6}$$

2) Exmaple of modeling data source relationship with BN: Troyanskaya *et al.* [11] introduced a framework for integrating multiple data types and microarray analysis methods called the MAGIC (Multisource Association of Genes by Integration of Clusters). The main component of MAGIC is a BN that describes the relationship between possible data types and their grouping. For example, "two-hybrid" data and "reconstructed complex" data, both of which infers physical contact between bio-molecules of interest, is linked to a group called "physical association". The BN structure and the prior probabilities that describe the relationship between multiple data types and microarray analysis methods is constructed based on expert knowledge. Given set of evidences (data types), the BN is used to compute the posterior probability about whether a pair of genes has a functional relationship.

3) Examples using BN to varify pair-wise variable relation with multiple data source: Jensen et al. [12] discussed the data integration of gene expression data, ChIP binding data, and promoter sequence data to infer whether there is a relationship between pairs consisting of a gene and a transcription factor (TF) in order to construct a biological regulatory network. The BN is used to model relationship between at two to three variables (a gene and one or two TFs) and between the three data types. A single posterior distribution for all unknown parameters are summarized in the following:

$$\begin{aligned} &p(C, w, \Theta|g, f, m, b) \propto \\ &\quad p(g|f, C, \Theta).p(C|m, b, w).p(\Theta, w) \end{aligned} \tag{7}$$

where Θ is a collection of linear model parameters, *p(g|f,C,*Θ*)* stands for first level with gene expression as a linear function of TF expression and *p(C|m,b,w)* is second level with chip binding data and promoter sequence data and *p(*Θ*,w)* is the prior distribution for TF-specific prior weight.

Xing et al. [13] applied BNs for genomic data integration to reduce the misclassification rate in Protein-Protein Interaction (PPI). They proposed a method called nonparametric Bayes ensemble learning (NBEL), which is a nonparametric approach that dynamic integration data type by automatically up-weighting informative data source and down-weighting less informative and biased sources. Pairs of proteins are evaluated on the different data sources to determine whether there is as relationship between them. The pairwise relations are then used to construct the PPI network. The posterior probability of an interaction in pair of proteins i on data Y and the distribution f is formulated as follows:

$$\begin{aligned} &Pr(z_i = 1|Y, f) = \\ &\frac{\Psi_i \prod_{j=1}^{p} f_{1j}(y_{ij}}{\Psi_i \prod_{j=1}^{p} f_{ij}(y_{ij}) + (1 - \Psi_i) \prod_{j=1}^{p} f_{0j}(y_{ij})} \end{aligned} \tag{8}$$

where Ψ_i is prior probability of interaction in pair i, the value for Ψ_i can defer for uninformative data source. f_{0j} is the unknown distribution of the j^{th} score across protein pairs that do not interact, and f_{1j} is the unknown distribution of the j^{th} score across protein pairs that do interact. Y denote an data matrix, with rows corresponding to different protein pairs and columns to different types of scores from different data sources, y$_{i1}$,$\cdots$,y$_{ip}$.

The models developed to address different problems took advantage of strong mathematical basis of BN formulation, its natural capability to handle uncertainty and robustness in handling small changes in the model. BN provides a suitable framework for combining highly heterogeneous experimental data with expert knowledge. Moreover, the model developed is relatively easy to interpret and understand. However, major challenges for these BNs are the scalability issues and how to better model the prior probability distributions of random variables.

III. Multiple Kernel Based Methods

A. Kernels

Kernel based methods are nonparametric learning methods that utilize kernel functions [14] to define implicit similarity between the pair of samples in the data according to variables that describe the data. There are several advantages of kernel based methods. One of the advantages is that no prior assumptions about the distribution of the data are needed due to nonparametric characteristic of kernels. In addition, kernel functions can be used to model non-linear relationship between variables. Furthermore, the size of kernels are dependent only on the sample sizes and not on the number of variable or features, which makes it ideal for high dimensional data where the number of features are large.

The three most common kernel functions used are linear (k_{Lin}; Eq.9), polynomial (k_{Poly}; Eq.10) and Gaussian (k_{Gauss}; Eq.11). Considering two data point x_i and x_j the three kernel functions are formulated as follows:

$$k_{Lin}(x_i, x_j) =< x_i.x_j >, \tag{9}$$

$$k_{Poly}(x_i, x_j) = (< x_i.x_j > +1)^q, \tag{10}$$

and

$$\begin{aligned} k_{Gauss}(x_i, x_j) = exp(-||x_i - x_j||_2^2/s^2), \\ s \in R_{++}. \end{aligned} \tag{11}$$

B. Multiple Kernel Learning

Multiple kernel (MK) learning methods are set of methods that utilize combinations of kernels in the machine learning process. First part of MK learning is kernel fusion where multiple kernels are combined to form one kernel matrix. MKs can be created using a single data type with varying kernel functions and parameters which are then combined (kernel fusion) to achieve better learning results [15]. For the purpose of integrative analysis of heterogeneous data, kernel can be created for each data types and combined to be used with various kernel based learning methods.

There are two major advantages of MK methods compared to BNs in integrative analysis. First, they are often easier to modeling with no prior variable relations need to be modeled or learned. Second, in MK methods, data types need not be normalization prior to integration. That is, each raw representation of data types with varying scale, dimension, and distribution is transformed to a feature space.

Second part of MK learning is performing various kernel base learning methods, such as kernel perceptron, support vector machines (SVM), support vector regression, and kernel principal components analysis. Among the kernel based learning methods, SVM is the most widely used learning method for classification. SVM maximizes the marginal distance in the feature space using the discriminant function $f(x) = (w.\phi(x)) + b$. This results in a quadratic optimization problem

$$min\frac{1}{2}||w||^2 + C\sum_{t=1}^{N}\xi_t \qquad (12)$$

with respect to w $\in$ R^s, $\xi \in$ R$_+^N$, b $\in$ R subject to

$$y_t(< w, \Theta(x_t) > +b) \geq 1 - \xi_t \qquad (13)$$

where *b* is the bias term, *w* is the vector of weight coefficients, *C* is a predefined positive trade-off parameter between model simplicity and classification error, ξ is the vector of slack variables.

The optimization problem is solved using Lagrangian dual function. Thus the discriminant function can be rewritten as the following equation:

$$f(x) = \sum_{t=1}^{N}\alpha_t y_t k(x_t, x) + b \qquad (14)$$

where *k*(x_t,*x*) denotes the kernel function.

In addition to classification, MK learning also relates to other learning tasks other then classification such as feature selection and distance metric learning [15]. In feature selection, multiple kernels are used for learning from heterogeneous data sources and nonlinear variable selection. A study reported for group Lasso [16], in which features are well ordered into groups, and selection is conducted at group level. In order to address issues like quadratic growth of the kernel matrix with respect to the data, an ensemble of the kernel was proposed [17] inspired by the ensemble and boosting methods. In distance metric learning [18], MK is used to as the metrics to find similarity between inter class instances.

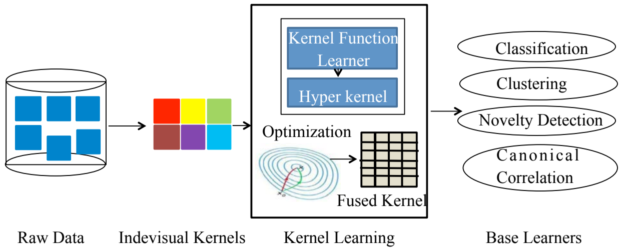

Fig. 3. Multiple kernel learning framework.

Fig. 3 represents the overall framework of multiple kernel learning approaches.

Gonen and Alpaydin [15] presented an in depth survey of MK learning methods focusing on kernel fusion methods categorized according to type of combination, functions optimized, and optimization approached used. According to their description, the basis of MK learning is an extension of the dual problem, which depends on the kernel description, whereas the heterogeneity deals with handling and converting data from different data structure into kernel matrices. The objective of the dual problem is to combine these kernel matrices such that the kernel coefficients optimize the overall objective, also known as kernel fusion.

Kernel fusion is an integrative part of MK learning framework. It facilitates the learning of an appropriate kernel to form an optimal kernel matrix. Kernel matrixes generated from individual data types can be combined linearly, nonlinearly, or data-dependently to maximize similarity between kernels, minimize the sum of regularization and error terms, or maximize the likelihood estimate using fixed rule, heuristics, optimizations, Bayesian, or boosting methods [15], [19], [20] in forming a fused kernel matrix.

Parameters in the MK learning can be learned in two stages or in one stage [15]. In the two-stage process, parameters associated with the kernel fusion are optimized and then they are used to learn the parameters of the base learners. In the one stage process, parameters of the kernel fusion and the base learner are optimized simultaneously.

The overall performance of MK learning depends on many factors, including number of kernels selected, training time, efficiency in terms of solution quality, base learner, and data set size as well as the choice of the kernel fusion method. The reported methods in the literature usually consider one or few factors for comparison and analyze the algorithm behavior.

C. Applications of Multiple Kernel Based Methods

Gene prioritization aims at identifying the significant causative gene in disease analysis. The main objective is to assign ranks among the gene based on their relevance to the biological process and select causative genes amongst the highest ranked genes. Recently in gene prioritization, multiple data sources such as gene expression, methylation, and mutation data, are integrative analysis to identify candidate genes most likely to be associated with or causative of a disorder. De Bie et al. [1] and Mordelet & Vert (ProDiGe) [21] used linear and weighted linear combination of kernels from multiple data sources to prioritize genes. They observed that kernel methods are accurate and robust against noise.

Multiple kernel learning constituted a powerful methodology as it allows integrating multiple data sets that extract non-linear features while representing variables. These models are more scalable and provide computational stability. However, compared to BNs base methods, theoretical results of such models are typically harder to prove.

IV. Conclusion

In this study, we have considered two different classes of integration analysis methods for heterogeneous data: Bayesian network (BN) based methods and multiple kernel (MK) based methods. Both methods provide an efficient means for integrating the data of different views of a subject. BN and MK can be selected based on the characteristics of the variables and data types of the problem. BNs are efficient when BN structures, i.e. relationship among variables and among the data views, can be model accurately. This is especially useful when existing knowledge about the variables and data types needs to be incorporated. On the contrary, MK is a better choice when knowledge about the variable relations is not explicitly known since normalization of each of the data types and prior assumption about the data distribution is not needed in MK modeling. Another advantage of BN includes easiness of providing theoretical proofs while it is difficult to theoretically prove the correctness of MK methods in prediction. However, BN requires both data normalization and knowledge of prior distribution, which can affect the result significantly if done incorrectly.

Acknowledgment

This project is supported by the NRF of Korea funded by the MSIP (2013005259) and by the MSIP, Korea, supervised by the NIPA (NIPA-2013-H0203-13-1001).

References

[1] T. De Bie, L. C. Tranchevent, L. M. van Oeffelen, and Y. Moreau, Kernel-based data fusion for gene prioritization, *Bioinformatics*, vol. 23, no. 13, pp. i125i132, 2007.

[2] K. Tsuda, S. Uda, T. Kin, and K. Asai, Minimizing the cross validation error to mix kernel matrices of heterogeneous biological data, *Neural Processing Letters*, vol. 19, no. 1, pp:6372, 2004.

[3] S. Qiu, and T. Lane, A framework for multiple kernel support vector regression and its applications to siRNA efficacy prediction, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 6, no. 2, pp:190199, 2009.

[4] N. Subrahmanya, and Y. C. Shin, Sparse Multiple Kernel Learning for Signal Processing Applications, *IEEE Transactions Pattern Analysis and Machine Intelligence*, vol. 32, no. 5, pp. 788-798, May 2010.

[5] S. S. Bucak, R. Jin, and A. K. Jain, Multiple Kernel Learning for Visual Object Recognition: A Review, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 7, pp. 1354-1369, July 2014.

[6] D. Rhodes, S. Tomlins, S. Varambally, V. Mahavisno, T. Barrett, S. Kalyana-Sundaram, D. Ghosh, A. Pandey, and A. Chinnaiyan, Probabilistic model of the human protein-protein interaction network, *Nature Biotechnology*, vol. 23, pp. 951959, 2005.

[7] C.J. Vaske, S.C. Benz, J.Z. Sanborn, D. Earl, C. Szeto, J. Zhu, J. M. Stuart, "Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM", *Bioinformatics*, vol. 26 no. 12, pp. i23745, 2010.

[8] O. Gevaert, F. De Smet, D. Timmerman, Y. Moreau, B. De Moor, "Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks." *Bioinformatics*, vol. 22 no. 14, pp. e184190, 2006.

[9] G. F. Cooper, and E. Herskovits, A Bayesian method for the induction of probabilistic networks from data, *Machine Learning*, vol. 9, pp.309347, 1992.

[10] D. Heckerman, A tutorial on learning with bayesian networks, Microsoft Research; 1995.

[11] O. G. Troyanskaya, K. Dolinski, A. B. Owen R. B. Altman, and D. Botstein, A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae), *PNAS*, vol. 100, no. 14, pp. 8348-8353, 2003.

[12] S. T. Jensen, G. Chen , and C. J. Stoeckert Jr., Bayesian variable selection and data integration for biological regulatory network, *The Annals of Applied Statistics*, vol. 1, no. 2, pp. 612633, 2007.

[13] C. Xing, D. B. Dunson, Bayesian Inference for Genomic Data Integration Reduces Misclassification Rate in Predicting Protein-Protein Interactions, *PLoS Computional Biology* 7(7): e1002110, 2011.

[14] J. Shawe-Taylor, and N. Cristianini, Kernel Methods for Pattern Analysis. Cambridge University Press. 2004.

[15] M. Gonen and E. Alpaydin, Multiple Kernel Learning Algorithms, *Journal of Machine Learning Research*, vol. 12, pp. 2211-2268, 2011.

[16] F. R. Bach, Consistency of the group Lasso and multiple kernel learning, *The Journal Machine Learning*, vol. 9, pp. 11791225, Jun. 2008.

[17] K. P. Bennett, M. Momma and M. J. Embrechts, MARK: A Boosting Algorithm for Heterogeneous Kernel Models, In *Proc. KDD-2002: Knowledge Discovery and Data Mining*, pp. 24-31, 2002.

[18] T. Hertz, Learning distance functions: Algorithms and applications, Ph.D. dissertation, Hebrew Univ. Jerusalem, Jerusalem, Israel, 2006.

[19] J. S. Hamida, C. M.T. Greenwood, and J. Beyene, Weighted kernel Fisher discriminant analysis for integrating heterogeneous data, *Computational Statistics and Data Analysis*, vol. 56, pp. 20312040, 2012.

[20] T. De Bie, L.C. Tranchevent, L. M. van Oeffelen, and Y. Moreau, Kernel-based data fusion for gene prioritization, *Bioinformatics*, vol. 23, no. 13, pp. i125-i132, Jul 2007.

[21] F. Mordelet, and J. P. Vert, ProDiGe: Prioritization Of Disease Genes with multitask machine learning from positive and unlabeled examples, *BMC Bioinformatics*, 12:389, 2011.