A Comparative Study of Hidden Markov Models Learned by Optimization Techniques using DNA data for Multiple Sequence Alignment

A.Priyanka, K.Sathiyakumari

Abstract— Efficient approach are based on probabilistic models, such as the Hidden Markov Models (HMMs), which currently represent one of the most popular techniques for multiple sequence alignment. In order to use an HMM method for MSA, one has to perform the parameter learning that is, to find the best set of state transition and output probabilities for an HMM with a given set of output sequences. In previous system, inspired by the free electron model in metal conductors placed in an external electric field here propose a novel variant of the PSO algorithm, called the random drift particle swarm optimization with diversity-guided search (RDPSO-DGS), and apply it to HMM training for MSA. In proposed system the two novel algorithms such that random drift firefly with diversity-guided search (RDFF- DGS) and random drift bat optimization with diversity-guided search (RDBO- DGS). It has fine adjustment of the parameters in this algorithm. In proposed algorithms are well effective than the existing system in terms of efficiency rate and computation cost of the system. That the HMMs learned by the RDFF and RDBO are able to generate better alignments. The experimental results show the RDBO-DGS gives the high accuracy 95% comparing to other algorithms for the herpes virus DNA data set it implement in MATLAB R2012a.. 100 data items are used for this research work, further can also use any Virus DNA for this alignment. It gives corresponding accuracy based on the data set. The remaining paper is organized as follows; Section 1. Describes introduction about the multiple sequence alignment. Section 2. Covers HIDDEN MARKOV MODELS FOR MSA. Section 3. Converse the related works behind in multiple sequence prediction. Section 4. Focus on experimental results comparison. Finally, Section 5. Discuss about the conclusion and feature work.

Index Terms— HMM; MSA; parameter learning; RDPSO; RDPSO-DGS, Random Drift firefly; Random Drift bat optimization.

1 INTRODUCTION

MULTIPLE sequence alignment (MSA) of nucleotides, or amino acids, is one of the most important and challenging problems in bioinformatics. It is an extension of a pairwise alignment to incorporate more than two sequences at a time. Multiple alignment methods try to align all of the sequences in a given query set, and the resulting aligned sequences are often used to construct phylogenetic trees, to find protein families, to predict secondary and tertiary structures of new sequences, and to demonstrate the homology between new sequences and existing families [1]. MSAs require more sophisticated methodologies than pairwise alignments since they are more computationally complex, and most formulations of the problem lead to NP-complete combinatorial optimization problems [2].

The dynamic programming method is theoretically applicable to any number of sequences. However, since it is computationally expensive in both time and memory, it is rarely used for more than three or four sequences in its most basic form [1]. One way to tackle this problem is to use a heuristic search, known as the “progressive alignment” technique, that builds up a final alignment by combining pairwise alignments, beginning with the most similar pair and progressing to the most distantly related ones [3], [4]. The most popular MSA program using progressive alignments is ClustalW, which is suitable for alignments of sequence sets with medium lengths [3]. T-Coffee is another common progressive alignment program, which is slower than ClustalW, but generally produces more accurate alignments for distantly related sequence sets [5]. ProbCons is a program for progressive alignment of protein sequences, i.e., a progressive protein MSA tool [6]. It incorporates the so-called probabilistic consistency technique, a novel scoring function for comparing multiple sequences.

Another set of approaches for MSA, known as the iterative methods, work similarly to progressive methods, but repeatedly align the initial sequences as well as adding new sequences to the growing MSA set. They produce multiple sequence alignments while reducing the errors inherent in progressive methods. MUSCLE is a popular iterative method and it improves on the progressive methods by using a more accurate distance measure to assess the relatedness of two sequences [7]. MAFFT is another popular MSA program, which incorporates different strategies including progressive methods (PartTree, FFT-NS-1, and LINS- 1), iterative methods (FFT-NS-i, L-INS-I, and G-INS-i) and structural alignment methods (Q-INS-I and X-INS-i) [8].

An alternative to progressive and iterative alignment methods is to employ stochastic optimization methods, such as the simulated annealing (SA) [9], [10] or evolutionary algorithms (EAs) [11], [12]. These approaches optimize an objective function, which measures the quality of multiple sequence alignment by updating the candidate alignments until an optimal alignment is found Other efficient approaches are based on probabilistic models, such as the Hidden Markov Models (HMMs), which currently represent one of the most popular techniques for multiple sequence alignment [13], [14], [15], [16]. There are several alignment programs in which variants of HMM based methods have been implemented. These programs are noted for their scalability and efficiency, although using an HMM method is more complex than using more common progressive methods. The most well-known HMM-based software packages for MSA are sequence alignment and modeling system (SAM) [17] and HMMER [18]. In order to use an HMM method for MSA, one has to perform the parameter learning, or the training task first; that is, to find the best set of state transition and output probabilities for an HMM with a given set of output sequences. The resulting HMM is then employed to create a sequence of gap insertions and deletion instructions to align the sequences. Generally, an HMM topology used for the MSA problem requires roughly as many states as the average length of the sequences in the problem. As such, one issue of the parameter learning in HMMs is that there is no known deterministic algorithm that can guarantee to yield an optimally learned HMM within reasonable computational time.

The most common way to deal with this problem is to employ an approximation algorithm based on statistics and reestimation. The Baum-Welch (BW) algorithm, known as the forward-backward algorithm, is the most widely used example of such algorithms [16]. The gradient methods [13] were also used to estimate the parameters of an HMM, but these methods are local search techniques that usually result in suboptimally trained HMMs. Another possibility is to estimate the parameters of an HMM by random optimization algorithms, such as the SA [19] and EAs [20]. For example, the well-known alignment software HMMER employs SA for HMM training. However, there always are complains that the SA and EAs encounter some problems, such as lack of local search ability, premature convergence and slow convergence speed. Particle swarm optimization (PSO) algorithm, a relatively recent population-based random search technique, has demonstrated its better performance than EAs and SA in HMM training for MSA.

This study focuses on HMM-based methods for MSA due to their scalability and efficiency. However, the number of states and parameters of an HMM increases with the average length of the sequences to be aligned, so that the dimensionality and complexity of the training problem of the HMM also increase with the average length. Although the HMM can easily be scaled to align large sequences efficiently, its training task is essentially a high-dimensional and multimodal optimization problem challenging to solve. Thus, our goal in this study is to develop an efficient global optimization method to train the HMM for MSA.

Applications of data mining to bioinformatics include gene finding, protein function domain detection, function motif detection, protein function inference, disease diagnosis, disease prognosis, disease treatment optimization, protein and gene interaction network reconstruction, data cleansing, and protein sub-cellular location prediction. For example, microarray technologies are used to predict a patient’s outcome. On the basis of patients’ genotypic microarray data, their survival time and risk of tumor metastasis or recurrence can be estimated. Machine learning can be used for peptide identification through mass spectroscopy. Correlation among fragment ions in a tandem mass spectrum is crucial in reducing stochastic mismatches for peptide identification by database searching. An efficient scoring algorithm that considers the correlative information in a tunable and comprehensive manner is highly desirable.

2 Hidden Markov Models For MSA

2.1 Topology of HMMs for MSA

For multiple sequence alignment, a Hidden Markov model denoted by λ consists of a set of q states (S1, S2, …, Sq) that are divided into three groups: match (M), insert (I) and delete (D) [15]. In addition, there are two special states, namely, the begin state and the end state.

States are connected to each other by transition probabilities a_{ij} where $0 \leq a_{ij} \leq 1\ (1 \leq i,j \leq q)$ and $\sum_{j=1}^{q} a_{ij} = 1 (1 \leq i \leq q)$. The delete states, the begin state and the end state do not emit observable symbols, so they are called silent states.

Starting from the begin state and until the end state, the HMM generates sequences, namely, strings of observable symbols, by making nondeterministic walks that randomly go from one state to another according to the transition probabilities. Each walk yields a path of visited states $\pi = (\pi_1, \pi_2, \dots, \pi_p)$ and a sequence consisting of emitted observable symbols on the path. When the HMM is applied to MSA, the sequence of observable symbols is given in the form of an unaligned sequence. The goal of multiple sequence alignment is thus to find a path π which generates the best alignment. It is possible to use the forward and Viterbi algorithms to determine,$P(o|\lambda)$, the probability of a given sequence o generated by the HMM λ, and derive the path π with maximal probability of generating the sequence o.

2.2 Learning HMMs for MSA

For a given sequence o and a hidden Markov Model λ, the goal of the learning task is to estimate the parameters, i.e., the transition and emission probabilities, of the HMM λ such that $P(o|\lambda)$ is maximized. The learning task is usually performed by either the Baum-Welch technique that is based on statistical re-estimation formulas or by random search methods such as the simulated annealing (SA) [19] or evolutionary algorithms (EAs) [20]. Before parameter estimation, the length of the HMM should be determined. A commonly used estimate is the average length of the sequences to be aligned. After parameter estimation, a better length of the HMM can be chosen by using a heuristic method known as the model surgery [15].

The quality of the HMM needs to be evaluated in the process of parameter estimation. Generally, a log-odds (LO) score is used for this purpose, which is based on a log-likelihood score.

Where O = ($O_1, O_2, \dots, O_N$) is the set of unaligned sequences, λ is the learned HMM, and $\lambda_{\mathbf{null}}$ is a null-hypothesis model. Here, a random model is chosen as the null-hypothesis model. The random model or the null-hypothesis model is the same for all the tested training algorithms for a given sequence set.

3 Related Works

Jullie D. Thompson, Desmond G. Higgins and Toby J. Gibson[4] progressive the multiole sequence alignment method. First, individual weights are assigned to each sequence . second, amino acid substitution matrices are veried. Third, residue-specific gap penalties and locally reduced gap in hydrophilic regions. Fourth, positions in early alignments where gaps have been opened received locally reduced gap penalties to encourage. where all of the sequences in a data set are very similar (e.g. no pair less than 35% identical), CLUSTAL W will find an alignment which is difficult to improve by eye. The algorithm greedily, iterative, progressive alignment algorithm, dynamic programming algorithm. Accuracy, first, the residue weight matrices, it improved by the accuracy of sequence alignment.

Da-Fei Feng and Russell F. Doolittle[3] describes the pair wise alignment algorithm iteratively to achieve the multiple alignment of a set of protein sequences and to construct an evolutionary tree depicting their relationship. The method has been applied to three sets of protein sequence: 7 superoxide dismutases, 11 globins, and 9 tyrosine kinase-like sequence. Multiple alignments and phylogenetic trees for these sets of sequences were determined and compared with trees derived by conventional pairwise treatments. The algorithm of Needleman and Wunsch(1970) was used in a three-matrix form(Fredman 1984).

Cedric Notredame, Desmond G. Higgins and Jaap Heringa[5] describes a new method (T-Coffee) for multiple sequence alignment that provide a dramatic improvement in accuracy. With T-Coffee we pre-process a data set of all pair-wise alignments between the sequences, high scoring segments that show consistency within the data set. Two main features. First, it provides a simple and flexible means of generating multiple alignments. The second main feature of T-Coffee is the optimization method. Used T-Coffee Algorithm also with genetic Algorithm and branch and bound algorithm. T-Coffee is 9.7% more accurate than the next-best method, Prrp.

Chuong B. Do, Mahathi S.P. Mahabhashyam, Michael Brudno, and Serafim Batzoglou[6] evaluate across a wide range of organisms, biologists need accurate tools for multiple sequence alignment of protein families. Its performance on several standard alignment benchmark data sets. On the BAliBASE, SABmark, and PREFAB benchmark alignment databases. The features of ProbCons that give it a strong increase in performance, we compared several ProbCons variants the “Twilight Zone” set from the SABmark alignment database. Viterbi algorithm, ProbCons algorithm. Viterbi algorithm gives maximum expected accuracy.

Robert C. Edgar[7] describe MUSCLE, a new computer program for creating multiple alignments of protein sequences. The MUSCLE program, source code and PREFAB test data are freely available at http://www.drive5.com/muscle. MUSCLE algorithm using here, the speed and accuracy of MUSCLE are compared with T-Coffee, MAFFT and CLUSTALW on four test sets of reference Alignments. MUSCLE achieves the highest, or joint highest, rank in accuracy on each of these sets.

Cedric Notredame and Desmond G. Higgins[12] describe a new approach to multiple sequence alignment using genetic algorithms and an associated software package called SAGA. Test cases chosen from Pascarella structural alignment data base and chymotrypsin sequences. Used SAGA algorithm. We compared these results with those given by CLUSTAL W on the same test cases. SAGA performs more accurately than CLUSTAL W on data sets of realistic size.

4 Experiment And Result

This research is mainly focus on predicting possibilities of Multiple Sequence Alignment using firefly and Bat optimization Algorithms.

4.1 Learning HMMs for MSA with RDPSO or RDPSO-DGS

For a given sequence set, when the RDPSO or RDPSO-DGS is used to perform HMM learning for multiple sequence alignment, the length of the HMM is kept constant during the learning process and only the parameters of the HMM are optimized. As indicated in Section 2, the length of the HMM is set to the average length of the sequences in the given sequence set. After the learning process, a better length of the HMM can be chosen by the model surgery. That means the length of HMM varies with the different sequence sets, but it would be fixed for a given sequence set. The parameters to be estimated include the transition and emission probabilities. Thus, a candidate solution represented by the position of a particle is a real-valued vector containing l transitions and m emission probabilities, which means that the dimension of the search space for the HMM training is D = l + m.

In the RDPSO with diversity-guided search, the diversity measure at the kth iteration, denoted as d_k , is given by the average euclidean distance from the particle’s current position to the centroid of the swarm, as in[35]. That is,

$$d_k = \frac{1}{L\,.\,A}\sum_{i=1}^{L}\sqrt{\sum_{j=1}^{D}(X_{i,k}^{j} - \bar{X}_{k}^{j})^2} \qquad (1)$$

Where $\bar{X}_k^j$ is given by $\bar{X}_k^j$=$(1/L)\sum_{i=1}^{L} X_{i,k}^j$. A denotes the length of the longest diagonal in the search space, and D is the dimensionality of the problem. A lower bound d_{low} is set for d_k to prevent the diversity from constantly decreasing on the course of the search. If the diversity measure d_k drops belowd_{low}, it will be controlled in such a way to increase it until it is larger thand_{low}. A mutation operation exerted on the global best particle $P_{g,k}$ play such a role of diversity increasing.

4.2 Learning HMMs for MSA with RDFF-DGS

The sky filled with the light of fireflies is a marvelous sight in the summer in the moderately temperature regions. There are near to two thousand firefly species, and most of them produce short and rhythmic flashes. The pattern observed for these flashes is unique for most of the times for a specific species. The rhythm of the flashes, rate of flashing and the amount of time for which the flashes are observed are together forming a kind of a pattern that attracts both the males and females to each other. Females of a species respond to individual pattern of the male of the same species.

We know that the intensity of light at a certain distance r from the light source conforms to the inverse square law. That is the intensity of the light I goes on decreasing as the distance r will increase in terms of I α 1/r2. Additionally, the air keeps absorbing the light which becomes weaker with the increase in the distance. These two factors when combined make most fireflies visible at a limited distance, normally to a few hundred meters at night, which is quite enough for fireflies to communicate with each other.

Now we can idealize some of the flashing characteristics of fireflies so as to develop firefly-inspired algorithms. Flashing characteristics of fireflies is used to develop firefly-inspired algorithm. Firefly Algorithm (FA or FFA) developed by Xin-She Yang at Cambridge University in 2007, use the following three idealized rules:

1. All the fireflies are unisex so it means that one firefly is attracted to other fireflies irrespective of their sex.

2. Attractiveness and brightness are proportional to each other, so for any two flashing fireflies, the one has lower bright will move towards the one which is brighter. Attractiveness and brightness both decrease as their distance increases. If there is no one brighter than other firefly, it will move randomly.

3. The brightness of a firefly is determined by the view of the objective function. In other words, the luminous intensity of fireflies depends on the environment of the objective function. For example, in maximization problem, luminosity can simple directly with the objective function value composition proportion. For other situation, luminosity can be defined into similar genetic algorithm of adaptive value function.

For a maximization problem, the brightness is simply proportional to the value of the objective function. Other forms of the brightness could be defined in an identical way to the fitness function in genetic algorithms.

4.3 Learning HMMs for MSA with RDBO-DGS

The bat algorithm is a new swarm intelligence optimization method, in which the search algorithm is inspired by social behavior of bats and the phenomenon of echolocation to sense distance. The algorithm exploits the so called echolocation of bats. This behavior can be used to formulate the new bat algorithm. Yang used three generalized rules for bat algorithms:

1. All bats use echolocation to sense distance, and they also guess the difference between food/prey and background barriers in some magical way.

2. Bats fly randomly with velocity vi at position xi with a fixed frequency fmin, varying wavelength λ and loudness A0 to search for prey. They can automatically adjust the wavelength (or frequency) of their emitted pulses and adjust the rate of pulse emission r $\in$ [0, 1], depending on the proximity of their target.

3. Although the loudness can vary in many ways, we assume that the loudness varies from a large (positive) A0 to a minimum constant value Amin.

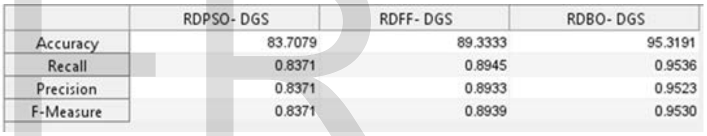
Bat algorithm bat behavior is captured into fitness function of problem to be solved. It consists of the following components:

- initialization
- generation of new solutions
- local search
- generation of a new solution by flying randomly
- Find the current best solution.

4.4 Data Collection

Multiple Sequence Alignment can be processed using the Protein/DNA datasets. In this works used Virus DNA data sets. From this link http://dna.cs.byu.edu/msa/. In this website having four different virus DNA, here used 100 data from herpes

TABLE 1
COMPARATIVE TABLE



	RDPSO- DGS	RDFF- DGS	RDBO- DGS
Accuracy	83.7079	89.3333	95.3191
Recall	0.8371	0.8945	0.9536
Precision	0.8371	0.8933	0.9523
F-Measure	0.8371	0.8939	0.9530

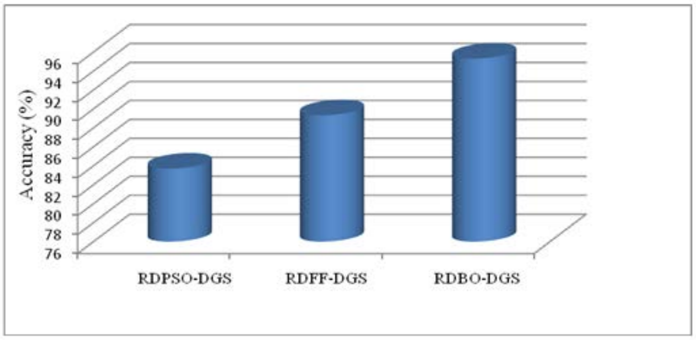


Fig..1 Comparison Accuracy chat

5 CONCLUSION

This paper focused on the HMM-based method for multiple sequence alignment. A novel firefly and Bat algorithm variant, the RDFF/RDBO algorithm along its improved version, named as the RDFF- DGS & RDBO- DGS, was proposed as a new optimization method and applied to train HMMs for MSA. The proposed RDFF-DGS & RDBO-DGS were used for the training of HMMs for MSA problems on herpes virus DNA data set. With the intension of improving the performance of the system as well as reduce the computation cost, proposing the two novel algorithms such that random drift firefly with diversity-guided search (RDFF- DGS) and random drift bat optimization with diversity-guided search (RDBO-DGS). It has fine adjustment of the parameters in this algorithm. Computation overhead of the multiple sequence alignment system is well reduced in this system. The experimental results show the RDBO-DGS gives the high accuracy 95% comparing to other algorithms for the herpes virus DNA data set. 100 data items are used for this research work. It gives corresponding accuracy based on the data set. Further it can also use any other type of Virus DNA data set/protein data set for the multiple sequence alignment. Also use two or three type of virus DNA to implement new type of implementation work.

REFERENCES

[1] D.W. Mount, Bioinformatics: Sequence and Genome Analysis. Cold Spring Harbor Laboratory Press, 2001.

[2] L. Wang and T. Jiang, "On the Complexity of Multiple Sequence Alignment," J. Computational Biology, vol. 1, no. 4, pp. 337-348, 1994.

[3] D.F. Feng and R.F. Doolittle, "Progressive Sequence Alignment as a Prerequisite to Correct Phylogenetic Trees," J. Molecular Evolution, vol. 25, no. 4, pp. 351-360, 1987.

[4] J.D. Thompson, D.G. Higgins, and T.J. Gibson, "CLUSTALW: Improving the Sensitivity of Progressive Multiple Sequence Alignment through Sequence Weighting, Position-Specific Gap Penalties and Weight Matrix Choice," Nucleotide Acids Research, vol. 22, no. 22, pp. 4673-4680, Nov. 1994.

[5] C. Notredame, D.G. Higgins, and J. Heringa, "T-Coffee: A Novel Method for Fast and Accurate Multiple Sequence Alignment," J. Molecular Biology, vol. 302, no. 1, pp. 205-217, Sept. 2000.

[6] C.B. Do, M.S. Mahabhashyam, M. Brudno, and S. Batzoglou, "ProbCons: Probabilistic Consistency-Based Multiple Sequence Alignment," Genome Research, vol. 15, no. 2, pp. 330-340, Feb. 2005.

[7] R.C. Edgar, "MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput," Nucleic Acids Research, vol. 32, no. 5, pp. 1792-1797, Mar. 2004.

[8] K. Katoh and D.M. Standley, "MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability," Molecular Biology and Evolution, vol. 30, no. 4, pp. 772- 780, Jan. 2013.

[9] J. Kim, S. Pramanik, and M.J. Chung, "Multiple Sequence Alignment Using Simulated Annealing," Bioinformatics, vol. 10, no. 4, pp. 419-426, July 1994.

[10] A.V. Lukashin, J. Engelbrecht, and S. Brunak, "Multiple Alignment Using Simulated Annealing: Branch Point Definition in Human mRNA Splicing," Nucleic Acids Research, vol. 20, no. 10, pp. 2511-2516, May 1992.

[11] K. Chellapilla and G.B. Fogel, "Multiple Sequence Alignment Using Evolutionary Programming," Proc. the First Congress on Evolution Composition, vol. 1, pp. 445-452, 1999.

[12] C. Notredame and D.G. Higgins, "SAGA: Sequence Alignment by Genetic Algorithm," Nucleic Acids Research, vol. 24, no. 8, pp. 1515-1524, Apr. 1996.

[13] P. Baldi, Y. Chauvin, T. Hunkapiller, and M.A. McClure, "Hidden Markov Models of Biological Primary Sequence Information," Proc. Nat'l Academy Sciences USA, vol. 91, no. 3, pp. 1059-1063, Feb. 1994.

[14] K. Karplus, C. Barrett, and R. Hughey, "Hidden Markov Models for Detecting Remote Protein Homologies," Bioinformatics, vol. 14, no. 10, pp. 846-856, 1998.

[15] A. Krogh, M. Brown, I.S. Mian, K. Sjolander, and D. Haussler, "Hidden Markov Models in Computational Biology: Applications to Protein Modeling," J. Molecular Biology, vol. 235, no. 5, pp. 1501- 1531, Feb. 1994.

[16] L.E. Baum, T. Petrie, G. Soules, and N. Weiss, "A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains," Annals of Math. Statistics, vol. 41, no. 1, pp. 164-171, 1970.

[17] R. Hughey and A. Krogh, "Hidden Markov Models for Sequence Analysis: Extension and Analysis of the Basic Method," Bioinformatics, vol. 12, no. 2, pp. 95-107, 1996.

[18] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison, Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge Univ. Press, 1998.

[19] S.R. Eddy, "Multiple Alignment Using Hidden Markov Models," Proc. the Int'l Conf. Intelligent Systems for Molecular Biology, vol. 3, pp. 114-120, 1995.

[20] S. Kwong, C. Chau, K. Man, and K. Tang, "Optimisation of HMM Topology and Its Model Parameters by Genetic Algorithm," Pattern Recognition, vol. 34, no. 2, pp. 509-522, Feb. 2001.

[21] L. Hubert and P. Arabie, "Comparing Partitions," *J. Classification,* vol. 2, no. 4, pp. 193-218, Apr. 1985. (Journal or magazine citation)

[22] R.J. Vidmar, "On the Use of Atmospheric Plasmas as Electromagnetic Reflectors," *IEEE Trans. Plasma Science*, vol. 21, no. 3, pp. 876-880, available at http://www.halcyon.com/pub/journals/21ps03-vidmar, Aug. 1992. (URL for Transaction, journal, or magzine)

[23] J.M.P. Martinez, R.B. Llavori, M.J.A. Cabo, and T.B. Pedersen, "Integrating Data Warehouses with Web Data: A Survey," *IEEE Trans. Knowledge and Data Eng.*, preprint, 21 Dec. 2007, doi:10.1109/TKDE.2007.190746.(PrePrint).