

Of Maps Bigger than the Empire*

Alberto Apostolico[†]

Purdue University and Università di Padova

Abstract

In a passage by J.L. Borges on the “exactitude of Science”, a fictitious author describes an Empire in which the art of Cartography “logró tal perfeccion que el mapa de una sola Provincia ocupaba toda la Ciudad, y el mapa del Imperio toda una Provincia”. With time, these huge maps wouldn’t be enough, and the Colleges of the Cartographers erected a map of the Empire that equalled in width the Empire itself... This paper concerns itself with increasing cases of pattern discovery and data mining in which synopses, indices and relationships thereof seem to grow faster and bigger than the phenomena they were meant to encapsulate. The paper then reviews specific examples of algorithmic and combinatorial constructs that proved capable of alleviating such paradoxes in the author’s recent work experience.

1. Introduction

In his “Viajes de Varones Prudentes¹” (Libro Cuarto, Cap. XLV, Lerida, 1658), a J. A. Suàrez

*Keynote for the 8-th International Symposium on String Processing and Information Retrieval - SPIRE 2001, Laguna de San Rafael, Chile, November 2001. Proceedings by IEEE Computer Society, to appear. Work supported in part by NSF Grant CCR-9700276, by NATO Grant CRG 900293, by Purdue Research Foundation Grant 690-1398-3145, and by the Italian Ministry of University and Research.

[†]Department of Computer Sciences, Purdue University, Computer Sciences Building, West Lafayette, IN 47907-1398, USA. axa@cs.purdue.edu and Dipartimento di Elettronica e Informatica, Università di Padova, Via Gradenigo 6/A, I-3131 Padova, Italy. axa@dei.unipd.it

¹The piece was written by Jorge Luis Borges and Adolfo Bioy Casares. English translation quoted from J. L. Borges, *A Universal History of Infamy*, Penguin Books, London, 1975: “... In that Empire, the craft of Cartography attained such perfection that the map of a single province covered the space of an entire City, and the Map of the Empire itself an entire Province. In the course of time, these extensive maps were found somehow wanting, and so the College of Cartographers evolved a Map of the Empire that was of the same scale as the Empire and that

Miranda narrated by J.L. Borges writes that “... En aquel Imperio, el Arte de la Cartografía logró tal perfección que el mapa de una sola Provincia ocupaba toda una Ciudad, y el mapa del Imperio, toda una Provincia. Con el tiempo, esos mapas desmesurados no satisficieron y los Colegios de Cartógrafos levantaron un Mapa del Imperio, que tenía el tamaño del Imperio y coincidía puntualmente con él. Menos adictas al Estudio de la Cartografía, las generaciones siguientes entendieron que ese dilatado Mapa era Inútil y no sin impiedad lo entregaron a las inclemencias del Sol y de los Inviernos. En los Desiertos del Oeste perduran despedazadas ruinas del Mapa, habitadas por animales y per mendigos; en todo el País no hay otra reliquia de las Disciplinas Geográficas.”

The increase in computation and communication power promises to reshape substantially the environment and operation of human activity. In the last decade alone, technological breakthroughs have led to widespread diffusion of personal computers, to the introduction of superscalar and fine-grain parallel machines of once unthinkable power and speed, in a process culminating for the time being in the recent announcement of the first teraflop (i.e., 10^{12} floating point operations per second) machine. Internet and WEB based technologies are only few years old, but have already revolutionized some deep and established sectors and patterns of activity. From only 1000 hosts connected in 1990, about 30 million were counted in Jan. 1998, connecting about 112 million users (70 in the U.S., 20 in Europe). The increase in connectivity and bandwidth of this medium will followup the transition from mainframe to personal computing leading into ubiquitous computing, a scenario within which, as it has been observed, humans are going to be shared among many computers.

coincided with it point for point. Less attentive to the study of Cartography, succeeding generations came to judge a map of such magnitude cumbersome, and, not without irreverence, they abandoned it to the rigours of Sun and Rain. In the Western Deserts, tattered fragments of the Map are still to be found, sheltering an occasional beast or beggar; in the whole Nation, no other relic is left of the Discipline of Geography.”

There is little doubt that the trend is towards greater use of electronic media, data bases and data repositories, within paradigms of activity characterized by ubiquitous processing of equally ubiquitous, intangible commodities. Thus, in the emerging information infrastructures, volumes of information will be amassed, disseminated and shared at an increasing pace. Data in the terabyte, even petabyte range (1 petabyte is 1 billion times 1 million characters) will pose unprecedented problems of management and policy. Effective access to, and manipulation of information will depend crucially on the efficiency with which information itself is structured, compressed, transmitted, stored and retrieved. Such issues subtend to a wide range of innovative applications, from Electronic Commerce and other web based activities to Music, Bioinformatics and more in general to scientific and commercial database development.

The nature and rate of these changes is quantitative on surface but will induce long qualitative leaps. In the forthcoming information flood, search for information without meaning will become impossible or useless (search engines still return thousands of Greek restaurants to a query about the "Parthenon"). In computer science jargon, we are moving from paradigms of search "by value" and search "by contents" to a new one of search "by meaning", a paradigm yet to be explored. To appreciate the difficulty this poses, it suffices to consider that already search by contents, which appears to be so easy in text, becomes quite difficult with other media such as pictures and sounds. The environment within which most of this processes must take place is like the famous Heraclitus' river: never twice the same. I shall argue next that even making all this data and information *accessible* leaves still an issue of making it *accessed*. Unless of course we wish to take the view that – to paraphrase the central problem of Episteme – Knowledge might exist out there, just like reality, with no need to reside once in a human mind.

2. Paradigms Old and New

The perception of problems centered on information dissemination, retrieval, and analysis is well rooted in the tradition of the computer and information sciences, where such facets have formed the subject of study for several decades. The focus of this paper, however, is on the completely new scenario and on the wild paradigm shift that are forced by the recent progresses of IC Technology. The new scenario is that data and information accumulate at a pace that makes it no longer fit for direct human inspection. The paradigm shift is that, in contrast to a primeval, persistent tenet of tra-

ditional information science and technology, the bottleneck in communication is no longer represented by the channel or medium but rather by the limited perceptual bandwidth of the final user: more and more often, the time and resources that need to be invested in order to gain access to information happen to be disproportionate to fruition time and value to the user, thereby defying the very purpose of access. A recent in-flight entertainment showed the happy purchaser of the latest satellite receiver as he spends the entire day pushing the next-channel button in search of what he would like to watch best. The message is clear: what good will it make being able to choose among 1000 movies when reading their summaries takes longer than watching one. We thus see that compared to traditional ICT, the challenge of maximizing the throughput to the final user has taken up entirely new meanings. With data being unendingly amassed, the prevailing problem is becoming one of how to limit and filter what a query shall return. Correspondingly, new compelling issues are faced of achieving effective synthetic descriptions, generating succinct characterizations, enhancing prominent features for the available data. Metadata buildup is an essential intermediate task towards these objectives. Along these lines, a whole new science and engineering of automated discovery will have to take shape, of which the grounds are just beginning to be laid. The mechanics of discovery, and scientific discovery at that, will undergo changes perhaps comparable to the scientific revolution itself. The implications brought about by such a dramatic change in perspectives have barely begun to be perceived.

At some core level in these endeavors, it comes also natural to identify the need for novel techniques supporting the automated discovery of patterns and their associations or "rules" in disparate contexts and media. The ability to automatically detect or generate patterns and associations will gradually become the only means of access to data and information too huge to be palatable. The techniques developed along these lines find ad hoc incarnations in diverse fields but also feature a distinctively unifying flavor.

The terms *pattern discovery* and *rule discovery* begun to be used recently in an attempt to encapsulate a repertoire of syntactic problems and tools akin to the identification of regularities such as repetitions, cadences, motifs, and joint or connected occurrences thereof in elementary discrete objects of various kinds. The natural predecessors of such problems and techniques were originally met as early as in the 70's in contexts such as, e.g., compiler design. Pattern and Rule Discovery may be regarded as a set of novel Pattern Matching problems brought about by the recent explo-

sive emergence of applications related to multimedia systems, computational molecular biology, very large database systems, worldwide information servers, and new software and hardware technology. These problems constitute a largely unexplored territory and pose both considerable opportunities and challenges.

Although the present discussion invests many distant and diverse areas, it helps to set the stage if we consider briefly the area of Bioinformatics. It is likely that many advanced discovery and mining techniques will find early incarnations in this area. Its development is motivated by the accumulation of data, especially nucleotide and protein sequences, too abundant and obscure to be examined and understood without the help of a computer. The data itself pertains to the complex pathways of information processing that take place in living organisms. Some of these information processing mechanisms can be regarded as information processing systems in their own merit, and are in fact susceptible of use as such like, e.g., in molecular computing. Thus there are two qualifications to Bioinformatics: it studies the application of computer and information sciences to the unveiling of the complex molecular structure of the living, and the information processing that takes place in living organisms. Due to the inherently scattered, cooperative nature of such gigantic endeavors as the Human Genome Project, Bioinformatics has recently found itself “de facto” scouting at the frontier of the information sciences at large, thereby leading in the way of network computing and other advanced forms of computation. This process is still under way. Irrespective of this, it is often projected that Bioinformatics will be to the 21st Century what Computer Science has been to the present one.

The collection of computational problems arising in contemporary molecular biology and genomics is large and growing. As WEB sites and data banks accumulate known proteins, DNA sequences, and 3D structures, increasingly fast and sophisticated discovery tools are sought. Antagonist trends represented by data explosion on one hand, and growing need for integration and cross-correlation on the other, makes advances in motif and rule discovery a foremost need in this domain. Manual search for relationships in the growing collection of data is no longer feasible. Capabilities akin to automatically cluster, classify, and annotate data across the traditional boundaries of individual databases require novel models and algorithms for the analysis and cross-annotation of scattered biological data. Bioinformatics continues thus to offer a stimulating micro-cosmos within which most models and problems can be studied in a relatively well defined environment and in somewhat controlled fashion. Sim-

ilar features are shared by many other environments, from Natural Languages to Speech Processing, from Music to Image Retrieval, and more.

3. Theories Bigger than Life

If the “dilatado Mapa” of Suárez Miranda looks so clearly “inútil”, what shall we say when confronted with maps that are in fact *bigger* than the empire. Some such phenomenon tends to manifest itself where there is accumulation of raw data, compound with heavy data processing that produces even more data. The risk is possibly largest at the crossing of disciplines that work by dissecting and classifying, notably, the Life Sciences, and disciplines that work by induction and synthesis such as Math, Statistics and Computer Science. The problem seems to reside with some misconception, that data processing necessarily leads to a compaction of the input data, whereas it might actually lead to explosion.

Because strings are such elementary, primeval structures it comes natural to look there first for examples. As it turns out, one does not need to look far. For practical reasons I will choose for illustration few cases taken from recent work in which I have been involved, but it should become apparent that such cases are abundant.

Statistical indices - The problem of characterizing and detecting over- or under-represented words in sequences arises ubiquitously in diverse applications and has been studied rather extensively in Computational Molecular Biology, where such patterns seem to be deeply implicated in various facets of gene regulation and function. In most approaches to the detection of unusually frequent words in sequences, the words (up to a certain length) are enumerated more or less exhaustively and individually checked in terms of observed and expected frequencies, variances, and scores of discrepancy and significance thereof. With biological sequences and whole genomes being increasingly amassed, tasks close to an exhaustive enumeration and storage of the entire subword statistics in a single archive become entirely conceivable. Still, such tasks pose daunting computational challenges, and their potential accessibility and use is far from obvious.

Assume we wanted to build a statistical table of all substrings of a string. How many entries should the table accommodate? We know that although there are $\Theta(2^n)$ distinct words of up to $n - 1$ characters, our string may only contain $O(n^2)$ distinct substrings, which corresponds to the roughly n ways to choose the starting position and as many ways to choose the end.

Thus, an exhaustive subword count table for a string, far from being a synopsis, is bigger than that string.

The situation does not improve if we restrict ourselves to statistically devious substrings. The steps typically taken in these cases are as follows. First, a model is postulated or derived for the source emitting the string, and also some measure of deviation is established: e.g., the difference between expected and counted occurrences, divided by some normalizing factor such as, e.g., the variance or its square root. With respect to this score, a threshold value is next chosen, so that a string is over or under-represented if its score exceeds the threshold in absolute value. With all that agreed upon, we introduce now an *observed* string for study and look for surprising substrings in it. The first thing to note is a striking asymmetry in the universe of surprises: since only $O(n^2)$ of the $\Theta(2^n)$ words can appear in our textstring, then most of the $\Theta(2^n)$ words are condemned to be surprisingly *under*-represented in any textstring since they are not there at all in the first place. If we now ask for how many words could be over-represented in the observed string, then based on our a-priori scores and threshold it may be that all $\Theta(n^2)$ of them are, say, over-represented (e.g., the observed string is formed by the two most probable symbols only). The fastest and most efficient program might thus give us an output we cannot use.

Tandem occurrences and association rules - The phenomenon under study escalates as we consider word aggregates and compounds. Among the problems in this class we find, for instance, the detection of all squares or palindromes in a string. Some of the optimal $O(n \log n)$ algorithms known for the detection of squares or tandem repeats make crucial use of a bound of $O(n \log n)$ on the output. It is not difficult to extend those algorithms to treat germane problems such as the discovery of pairs of occurrences, within a given distance, of a same string, or of a string and its reverse, and so on.

Consider now the problem of detecting repetitive phenomena that consist of unusually frequent *tandem* occurrences, within a pre-assigned distance in a string, of two distinct but otherwise unspecified substrings. By the two strings occurring in tandem, we mean that there is no intermediate occurrence of either one in between. Formally, the problem is as follows. Let x be a string of n symbols over some alphabet Σ and d some fixed non-negative integer. For any pair (y, z) of subwords of x , their *tandem index* $I(y, z)$ relative to x is the number of times that z has a closest occurrence in x within a distance of d from a corresponding, closest occurrence of y . The problem is then to find pairs of

subwords with surprisingly high tandem index.

The problem can be seen as one of association or rule discovery. As is well known, while traditional data base queries aim at retrieving records based on their isolated contents, here the focus is on the identification of patterns occurring across records in large collection of data. An *association rule* is an expression of the form $S_1 \rightarrow S_2$ where S_1 and S_2 are sets of data attributes endowed with sufficient *confidence* and *support*. Sufficient support for a rule is achieved if the number of records whose attributes include $S_1 \cup S_2$ is at least equal to some pre-set minimum value. Confidence is measured instead in terms of the ratio of records having $S_1 \cup S_2$ over those having S_1 , and is considered sufficient if this ratio meets or exceeds a pre-set minimum. Clearly, a statistic of the number of records endowed with the given attributes must be computed as a preliminary step, and this is often a bottleneck for the process of information extraction.

Back to our problem, we observe that, in principle, there might be $\Theta(n^4)$ distinct pairings of subwords of in x , where again the number of interesting or surprising associations in an *observed* textstring may far out-size the string itself. We could think of many other possible examples in this class, for instance, in connection with hyperlink extraction and setting in hypertext collections.

Markov source modeling - Probabilistic models of various classes of sources are developed in the context of coding and compression as well as in machine learning and classification. In the first domain, the repetitive structures of substrings are regarded as redundancies and sought to be removed. In the second, repeated subpatterns are unveiled as carriers of information and structure. Source modeling is made hard in practice by the fact that we do not know the source probabilities, the latter being actually rather fictitious entities or models. In fact, one pervasive problem is that of learning or estimating these probabilities from the observed strings. In summary, the problem is twofold. From an information theoretic standpoint, the question is how to define a notion of information relative to a class of sources. Once one such characterization is agreed upon, interesting algorithmic questions revolve around the computational cost inherent to the process of learning or estimating probabilities within that class.

Some popular probabilistic automata typically built in these contexts are subtended by uniform, fixed-memory Markov models. In practice, such automata tend to be unnecessarily bulky and computationally imposing both during their synthesis and use. One of

rences, then the number of equivalence classes is linear in the textstring. For any probabilistic model, it is enough then to list for each class (i.e., node in either the subword tree or DAWG) the score of the most devious word in that class. Any other word and score is either implicit in these representatives, or uninteresting, or both. In fact it is possible to set up ad-hoc hybrid structures for this purpose, which are even more succinct.

The combinatorial property intrinsic to Fact 4.1 can alleviate somewhat also the problem of association discovery discussed earlier. In fact, it is enough to consider a subset of pairs containing only n^2 pairs, i.e., those formed by pairing up classes corresponding to nodes in either the subword tree or DAWG. For any pair (w', z') neglected as a result of this, there is a pair (w, z) in the chosen set of pairs such that w' is a prefix of w and z' is a prefix of z , and the tandem index of (w', z') equals that of (w, z) . However, the resulting theory is still bigger than life in this case.

Some shadow of Fact 4.1 extends to the synthesis of the structure of uniform, fixed-memory Markov models. As mentioned, for sequences in important families, such as those arising in applications that range from natural language, to speech, handwriting, and molecular sequence analysis, the autocorrelation or “memory” exhibited decays exponentially fast with length. In other words, there is a maximum length L of the recent history of a sequence, above which the empirical probability distribution of next symbol given the last $L' > L$ symbols does not change appreciably. It is possible and customary to model these sources by Markov chains of order L , this maximum useful memory length. Even so, the exponential growth in size by such automata makes them rapidly unpractical. In recent work by Ron *et al* much more compact, tree-shaped variants of probabilistic automata (called *PSTs*) are built which assume an underlying Markov process of variable memory length not exceeding some maximum L . These variants were subsequently adapted and applied successfully to learning and prediction of protein families. The probability distribution generated by these automata is equivalent to that of a Markov chain of order L , but the description of the automaton itself is much more succinct. The process of learning the automaton from a given training set S of sequences requires $\Theta(Ln^2)$ worst-case time, where n is the total length of the sequences in S and L is the length of a longest substring of S to be considered for a candidate state in the automaton. Once the automaton is built, predicting the likelihood of a query sequence of m characters may cost time $\Theta(m^2)$ in the worst case.

In recent work with G. Bejerano, we have introduced automata equivalent to *PSTs* but having the desirable properties that their construction and use takes linear time. That is to say, in particular, that the size of the learned classifier does not exceed that of the observation upon which it is based. The crux in the improvement resides in speeding up a test that asks, virtually on all substrings s of the source string, whether there is a symbol $\sigma \in \Sigma$ such that:

$$\frac{\tilde{P}(\sigma|s)}{\tilde{P}(\sigma|suffix(s))} \geq r \quad \text{or} \quad \frac{\tilde{P}(\sigma|s)}{\tilde{P}(\sigma|suffix(s))} \leq 1/r,$$

where \tilde{P} denotes empirical probabilities or frequencies and $r \geq 1$ is a fixed parameter value. Essentially, it is possible to set up an algorithm to answer the collection of all those tests for all substrings of a textstring x in overall linear time and space. The main intuition is that, for subwords s and $s' = s\sigma$ in a same equivalence class of \equiv_x , the empirical conditional probabilities in the form $\tilde{P}(\sigma|s)$ must be 1 and thus there is no need to compute it explicitly.

A prudent choice in the identification of primitive patterns and constructs may lead to even more dramatic savings. Recently, a class of motifs called “irredundant” has been characterized having the property, that they can only grow linearly, as opposed to exponentially, with the size of the host string. This opens the way to the quest for fast and efficient methods of extraction for such motifs.

Given a string x on alphabet Σ , a string s on $\Sigma \cup \cdot$ such that the first and last characters are solid characters is a *motif* with location list $\mathcal{L}_s = (l_1, l_2, \dots, l_p)$ provided that p is maximal for s . The location list is thus an integral part of the motif, so that two distinct location lists must refer to distinct motifs. An important attribute of maximality is naturally associated with motifs. In intuitive terms, a motif s is *maximal* if we cannot make it more specific or longer while retaining the list \mathcal{L}_s of its occurrences in x up to a uniform offset of \mathcal{L}_s . Thus the only solid maximal motif with one occurrence is x itself, and for any maximal motif with at least one non-solid character we have $p \geq 2$. For instance, in $x = abcdabbd$, $s_1 = a..d$ with $\mathcal{L}_{s_1} = \{1, 5\}$, and $s_2 = b.d$ with $\mathcal{L}_{s_2} = \{2, 6\}$ are non-maximal motifs, while $s_3 = ab.d$ with $\mathcal{L}_{s_3} = \{1, 5\}$ is maximal. The notion of maximality hinges on two “saturation” conditions, in the sense that a maximal motif is maximal in composition, and maximal in length. However, the notion of maximality alone does not suffice to bound the number of different motifs. It can be shown that there are strings that have an unusually large number

of maximal motifs without conveying extra information about the input.

A maximal motif s is *irredundant* if the list \mathcal{L}_s of its occurrences cannot be deduced by the union of a number of lists of other maximal motifs. Conversely, we call a motif s *redundant* if s (and its location list \mathcal{L}_s) can be deduced from the other motifs *without* studying the input string s . More formally, a maximal motif s , with location list \mathcal{L}_s , is *redundant* if there exist maximal motifs s_i , $1 \leq i \leq p$, such that $\mathcal{L}_s = \mathcal{L}_{s_1} \cup \mathcal{L}_{s_2} \dots \cup \mathcal{L}_{s_p}$.

Redundancy is supported by some notion of motif combination. Two motifs s_1 and s_2 such that the solid characters “clash” cannot be combined. Thus, $s_1 = ab.d$ and $s_2 = acdd$ cannot be combined since $b \neq c$. However, if $s_1 = ab.d$ and $s_2 = a.dd$, then $s = a..d$ and $\mathcal{L}_s = \mathcal{L}_{s_1} \cup \mathcal{L}_{s_2}$.

We use \mathcal{B} to denote the set of irredundant motifs in a string x . Set \mathcal{B} is called the *basis* for the motifs of x . If we consider the set \mathcal{M} of all maximal motifs in x , \mathcal{B} is such that: (1) for each $s \in \mathcal{B}$, s is irredundant with respect to $\mathcal{B} - \{s\}$, and, (2) let $\mathbf{G}(\mathcal{X})$ be the set of all the redundant maximal motifs generated by the set of motifs \mathcal{X} , then $\mathcal{M} = \mathbf{G}(\mathcal{B})$. In general, $|\mathcal{M}| = \Omega(2^n)$, but a remarkable result by L. Parida *et al* shows that every string x of n symbols has a basis \mathcal{B} or set of irredundant motifs, such that $|\mathcal{B}| \leq 3n$. Hence it is always possible to account for and describe the rich combinatorics of the motifs in a string through the compact description set forth by a reasonably small basis.

5. Conclusion, Computational Theories of Surprise

I have tried to examine the reflections, within the small boundaries of my own recent work experience, of phenomena that seem to reach farther in our disciplines both with their origins and implications. Just as Stone Age did not end for lack of stones, the Information Age will not end for lack of Information. On the contrary, with data accumulating at the current pace, there is no guessing what the next big flood will look like. Classical Information Theory dealt with bringing as much information as possible to the end user. In this paradigm, Information is *surprise*, and the bottleneck is the Medium. Some notion of surprise seems solidly implicated in current and future managements of Information as well. However, there is increasing risk that the overhead of access or representation defy usefulness in this case. In the Information Theory to come the bottleneck is no longer the Medium, but rather the perceptual bandwidth of the Receiver. Under such conditions, the main task seems to be rather that of “reducing information and increasing insight”, as A.

Dress nicely put it. Setting up such an “Insight Theory” is likely to involve the predisposition of a number of novel computational models and implements. These, in turn, are likely to be centered on the optimization of parameters akin to compression ratios and more subtle, qualitative measures of input/output or throughput relationships. These parameters might even supplant in part more traditional indicators of computational complexity: of the several theoretically untractable problems in Computational Biology (e.g., multiple sequence alignment or protein folding) none seemed to represent good example in the perspective of this paper. This means that coping with such a scenario might be far more challenging than we were ever used to see.

Acknowledgements. I am indebted to Gustavo Stolovitzky for pointing out to me the passage by J.L. Borges, and to the Bioinformatics and Pattern Discovery Group at IBM T.J. Watson Center for the hospitality I enjoyed there during the Fall of 2000. I also wish to thank Mary Ellen Bock, Rich De Millo, Andreas Dress, Titti Guerra, Mike Waterman, for discussions and insights on the topic of this paper. Finally, I am grateful to my mother the late Rosita Alfano, born in Mendoza, Argentina, for entitling me to a little extra indulgence whenever I step on Philosophy while in Latin America.

References

- [1] Abe, N. and M. Warmuth. On the Computational Complexity of Approximating Distributions by Probabilistic Automata, *Machine Learning*, **9**, 205–260 (1992).
- [2] Agrawal, R., T. Imielinski, A.N. Swami. Mining Association Rules between Sets of Items in Large Databases. *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, Washington, D.C., May 26–28, 1993, pp. 207–216. ACM Press (1993).
- [3] Apostolico, A. Notes on Learning Probabilistic Automata, *Proceedings of DCC 2000*, Snowbird, IEEE Press (2000).
- [4] Apostolico, A. and G. Bejerano. Optimal Amnesic Probabilistic Automata or How to Learn and Classify Proteins in Linear Time and Space, *Proceedings of RECOMB2000*, Tokyo, April 2000 (1999) and *Journal of Computational Biology*, 7(3/4):381–393 (2000).

- [5] Apostolico, A., M. E. Bock, and S. Lonardi. Linear Global Detectors of Redundant and Rare Substrings. In J. A. Storer and M. Cohn, editors, *Proceedings of Data Compression Conference*, 168–177, Snowbird, Utah, (April 1999).
- [6] Apostolico, A., M. E. Bock, S. Lonardi and X. Xu. Efficient Detection of Unusual Words. *Journal of Computational Biology*, 7(1/2):71–94 (2000).
- [7] Apostolico, A., and Z. Galil (Eds.), *Pattern Matching Algorithms*, Oxford University Press, New York (1997).
- [8] Apostolico, A. and G. Satta. “Optimal Discovery of Subword Associations in Strings ”, Purdue CS-TR 99-042, submitted (1999).
- [9] Arikawa, S. and K. Furukawa (eds.). *Discovery Science*, Springer-Verlag LNAI 1721 (1999).
- [10] Barbara, D., W. DuMouchel, C. Faloutsos, P.J. Haas, J.M. Hellerstein, Y. Ionnidis, H.V. Jagadish, T. Johnson, R. Ng, V. Poosala, K.A. Ross, K.C. Sevcik. *The New Jersey Data Reduction Report, Bulletin of the Technical Committee on Data Engineering* 20, Vol. 4, pp. 3-45 (1997).
- [11] Bejerano, G. and G. Yona. Modeling Protein Families Using Probabilistic Suffix Trees. *Proceedings of RECOMB99* (S. Istrail, P. Pevzner and M. Waterman, eds.), 15–24, Lyon, France, ACM Press (April 1999).
- [12] Berry, M.J.A. and G.Linoff. *Data Mining Techniques*, Wiley (1997).
- [13] Blumer, A., J. Blumer, A. Ehrenfeucht, D. Hausler, M.T. Chen and J. Seiferas. The Smallest Automaton Recognizing the Subwords of a Text, *Theoretical Computer Science* , **40**, 31-55 (1985).
- [14] Brazma, A., I.Jonassen, I.Eidhammer, D.Gilbert. Approaches to the Automatic Discovery of Patterns in Biosequences. *Journal of Computational Biology* 5, pp. 279-305 (1998).
- [15] Das, G., R. Fleischer, L. Gąsieniek, D. Gunopulos, J. Kärkkäinen. Episode Matching, *CPM’97, Proceedings of the 8th Annual Symposium on Combinatorial Pattern Matching*, (A. Apostolico and J. Hein, Eds.), Springer Verlag LNCS **1264**, 12-27 (1997).
- [16] Duda, R.O., P.E.Hart, D.G. Stork. *Pattern Classification*, Wiley (2000).
- [17] Durbin, R., S. Eddy, A. Krogh, G. Mitchison. *Biological Sequence Analysis*, Cambridge University Press (1998).
- [18] Forchhammet, S., and J. Rissanen. Coding with Partially Hidden Markov Models, *Proceedings of the IEEE Data Compression Conference, DCC95*, 92–101, Snowbird, Utah (1995).
- [19] Mannila, H., H. Toivonen and A.I. Vercamo. Discovering Frequent Episodes in Sequences, *KDD’95, Proceedings of the 1st International Conference on Knowledge Discovery and Data Mining*, AAAI Press, 210-215 (1995).
- [20] Parida, L., I. Rigoutsos and D. Platt. An Output-sensitive Flexible Pattern Discovery Algorithm, *Proceedings of the 12th Annual Symposium on Combinatorial Pattern Matching, CPM01*, Jerusalem, July 1-3, 2001, Springer Verlag Lecture Notes in Computer Science (2001).
- [21] Piatesky-Shapiro, G. and W.J. Frawley, Eds. *Knowledge Discovery in Databases*. AAAI Press/MIT Press (1991).
- [22] Rigoutsos, I. , A. Floratos, L. Parida, Y. Gao, D. Platt. The Emergence of Pattern Discovery Techniques in Computational Biology, *Journal of Metabolic Engineering*, to appear (2000).
- [23] Rissanen, J. A universal Data Compression System, *IEEE Trans. Inform. Theory* **29**(5): 656–664 (1983).
- [24] Rissanen, J. Complexity of Strings in the Class of Markov Sources, *IEEE Trans. Inform. Theory* **32**(4): 526–532 (1986).
- [25] Ron, D., Y. Singer and N. Tishby. The Power of Amnesia: Learning Probabilistic Automata with Variable Memory Length. *Machine Learning*, 25:117–150 (1996).
- [26] Ukkonen, E. On-line Construction of Suffix Trees. *Algorithmica*, 14(3):249–260 (1995).
- [27] Waterman, M. *Introduction to Computational Biology: Maps Sequences and Genomes*, Chapman and Hall (1995).
- [28] Witten, I.H., A. Moffat, T.C. Bell. *Managing Gigabytes: Compressing and Indexing Documents*, Morgan Kauffman (1999).