4th International Conference "Neural Networks and Expert Systems in Medicine and Healthcare" NNESMED 2001
20-22 June 2001, Milos Island, Greece

Using Hidden Markov Models to Characterize Disease Trajectories

M. Ohlsson and C. Peterson

Complex Systems Division, Department of Theoretical Physics
Lund University, Lund, Sweden

M. Dictor
Department of Pathology
Lund University, Lund, Sweden

ABSTRACT: A novel approach is developed for predicting body trajectories for cancer progression, where conditional probabilities of clinical data are modeled using Hidden Markov Model techniques. Basically, each potential body site is encoded by an N-letter code, and a disease trajectory is described in terms of a string of letters. Patient data base records are then represented by such strings with different lengths, start points and end points. The approach is explored using pathology data for non-Hodgkin lymphoma augmented with an artificial data base generated according to observed distributions in the clinical data. For the Hidden Markov Models a Bayesian approach is taken using the Hybrid Monte Carlo method, producing an ensemble of models rather than a single one. Using a test set consisting of both real and random trajectories, we estimate the performance of our Hidden Markov Model models and also extract most probable profiles. Given the limited data set size the results are very encouraging.

INTRODUCTION

Predicting the anatomical pattern of involvement of diseases such as cancer is important for prognosis, therapy and clinical follow-up. In principle, such trajectories represent time series. However, data sampling in terms of patient checkups and measurements are typically highly non-regular and incomplete. Furthermore, a specific disease need not originate in the same anatomical site in all patients. Hence standard time-series regression tools cannot be employed, and one has therefore to compute from data, conditional probabilities for the series of events. The predicted power of such an approach might be limited for trajectories that occur infrequently. This requires building internal models of the data. We have pursued such an approach by using Hidden Markov Model (HMM) techniques [1] for modeling the probabilities for disease trajectories. This is done by encoding each potential body site by an N-letter code, e.g. A,B,C,...,N, and describing a disease trajectory in terms of a string of letters, e.g. ABBDEEKLNA. Patient data base records are then represented by such strings with different lengths, start points and end points, for example:

GMMGLMMNN
MEMGA
LLLLMMM
AIAIA
JJJJLLNDD
MHHAHAG

The situation is very similar to the one occurring in multisequence alignment when comparing DNA (4-letter code) or amino acid (20-letter code) strings within bioinformatics. In that case one allows for deletions and insertions (presumably originating during evolution), when comparing many strings to obtain a "consensus string". In our case deletion and insertion correspond to the absence and irregularity of clinical measurements; yet there should exist an underlying "consensus string" in terms of a sequence of model probabilities corresponding to the natural progression of the disease.

MATERIAL

We explore the approach using pathology data for non-Hodgkin lymphoma with a reduced site code list (14 single-letter codes) for describing disease progression [2]. In Table 1 the reduced site codes and their corresponding physical locations are given.

Table 1: The reduced site codes and the corresponding locations.

Site Code	Location
A (1)	Miscellaneous / unclear
B (2)	Bone
C (3)	Stomach
D (4)	Hollow viscus incl. repro organs
E (5)	Head, nose and mucosa
F (6)	Mediastinum
G (7)	Solid organ incl. repro organs
H (8)	Skin
I (9)	Subcutis / soft tissues
J (10)	Serosa
K (11)	Central nervous system
L (12)	Lymphoid, infradiaphragmatic
M (13)	Lymphoid, supradiaphragmatic
N (14)	Marrow

The database contains 6938 entries on 2652 patients and although limited in size, it is representative of the noisy and incomplete real-world conditions we intend to characterize. The distribution of record lengths and the number of entries for each

letter are shown in Fig. 1. In order to fully test our algorithm we have created an additional database with artificial data, generated according to the distribution of events observed in the clinical data.

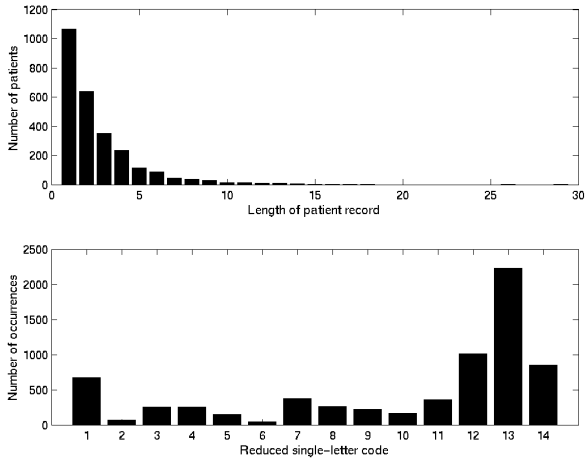

Figure 1: The distribution of patient record lengths in the database (upper graph) and the number of occurrences for each single letter code (lower graph).

Constraining the record lengths to the interval 5-15, the final data set consists of 359 patient records (sequence of reduced site codes). Shorter records than 5 turn out to be too difficult for the HMM to handle. This data set is then divided into a training set of 320 records and a test set of 39. An additional set of 39 records from the artificial data set are added to the test set in order to estimate the performance.

METHODS

In order to model the different code strings, representing the disease trajectories and allowing for gaps corresponding to non-existing measurements, a standard HMM architecture is used with a set of main, delete and insert nodes (see Fig. 2). A model of the data is then given by the transition and emission probabilities between the nodes (arrows). We use standard Dirichlet priors [1] to parameterize these probabilities. Rather than optimizing parameters once given the data, i.e. finding a single model, we generate an ensemble of models within a Bayesian framework. In other words, distributions of parameter values are obtained rather than a single set. For this calibration procedure one needs an efficient sampling scheme. As is frequently and successfully often used in the context of Bayesian neural network modeling [3], we employ the Hybrid Monte Carlo (HMC) method [4]. Loosely speaking, in this scheme Newtonian motions in parameter space (trajectories) are mixed with Metropolis jumps. The algorithmic parameters, trajectory lengths, step sizes and Metropolis steps are set to 7, 0.02 and 5000 respectively.

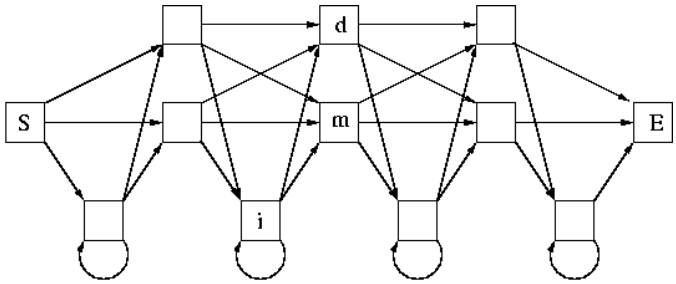

Figure 2: The standard HMM architecture used in this paper. **S** and **E** denote start and end state, respectively. The delete, main and insert states are marked as **d**, **m**, and **i**.

The length of the HMMs, M, is set to an estimated average record length of 10. The number of parameters in the HMM, with an N-letter code is given by (2M+1)N+9M+3, which for M=10 and N=14 is 387.

RESULTS

We explore the method in two ways. First we estimate the performance of the HMMs, by computing log probabilities, using parts of the real data set for training and for testing using both real and artificial data sets. Second, we compute profiles. Since the real data set is somewhat insufficient for the latter purpose, these calculations are entirely based upon the artificial data set.

Log probabilities: Once calibration is completed we are able to compute scores (log probabilities) for new trajectories, with a high score indicating that the trajectory is similar to those that were used to calibrate the HMM. Using the test set that consists of both real and random trajectories, we can estimate the performance of our HMMs. Figure 3 shows the distribution of the scores for the real and artificial test records using 100 HMMs sampled from the HMC procedure. The significant difference implies that the HMMs have learned the properties of the patient records. The average and the variance of the scores are shown in Table 2.

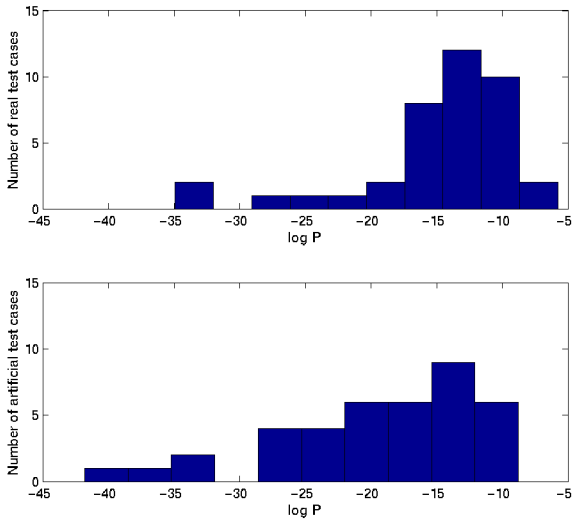

Figure 3: The distribution of the scores for the real test records (upper graph) and the artificial test records (lower graph).

Table 2: The average scores for the real and artificial test records.

	Mean score	Variance
Real test records	-15	6
Artificial test records	-20	8

Profiles: Next we test how well the method can recover profiles. To this end we generate artificial data sets corresponding to three artificial patient profiles. This data set reflects the real patient records in terms of length variation and noise levels. In Fig. 4 we show 12 randomly chosen records from the artificial dataset, which in total contain 700 records. The lengths vary in the interval 10-20 with an average length of 15. The number of site codes is 14, as for the real data set.

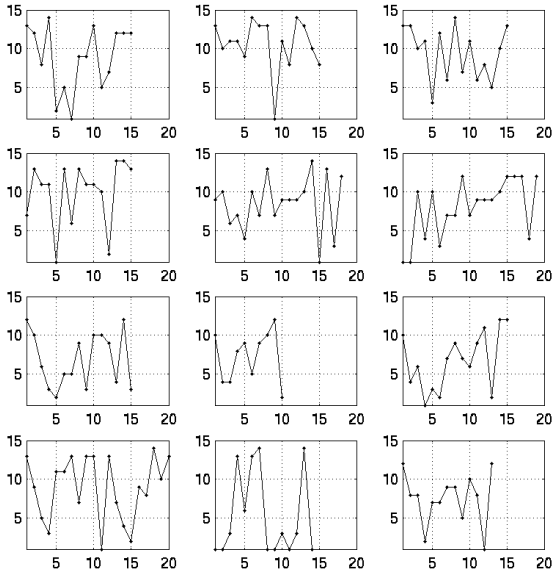

Figure 4: Site codes as functions of site positions for 12 randomly chosen records from the artificial data set.

After calibration with the HMC procedure, one can compute the most probable emissions along the main states for the HMM. In Fig. 5 the three artificial profiles are shown in the upper graph and the corresponding estimated profile in the lower graph. As can be seen this profile is almost identical to one of the artificial ones, which indicates the power of the method.

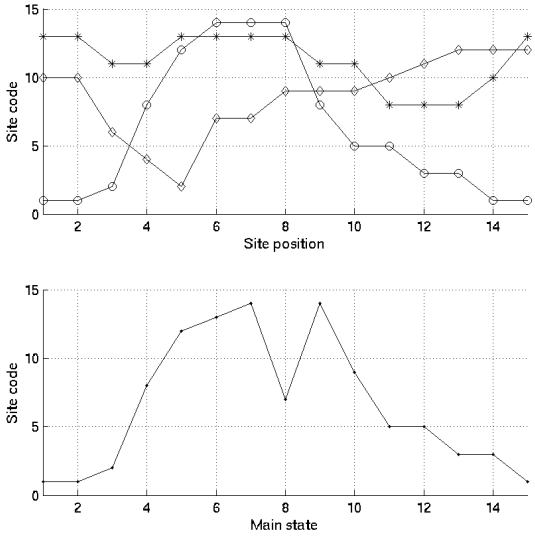

Figure 5: The three artificial profiles used when generating the records in the artificial data set (upper graph). The profile found by taking the most probable emissions along the main state of the HMM (lower graph).

SUMMARY

A Hidden Markov Model approach is applied to model patient disease trajectories using clinical data for non-Hodgkin lymphoma and artificial data created from clinical data distributions. The results look very promising given the limited statistics:

- Calibrated HMMs are used predict the behavior of clinical test sets with significance when compared to corresponding random data.
- HMMs calibrated with artificial data are able to correctly extract disease trajectory profiles.

The novel application of this methodology can be applied in health care situations far beyond the pathology database used as an example here. For example, one could include in the string codes for various types of treatment.

ACKNOWLEDGEMENTS

We are indebted to Anders Irbäck for providing us with an initial version of the HMM code used. This work was in part supported by the Swedish Foundation for Strategic Research.

REFERENCES

1. See e.g.. Durbin R., Eddy S.R., Krogh A., and Mitchison G.J., *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge UK, Cambridge University Press, (1998).
2. Dictor M., The surgical pathologist in a client/server computer network: work support, quality assurance and the graphical user interface. *Mod. Pathol*. **10,** 259-266 (1997).
3. Neal R.M., *Bayesian Learning for Neural Networks*. Lecture Notes in Statistics No 118, New York, Springer-Verlag (1996).
4. Duane S., Kennedy A.D., Pendleton B.J., and Roweth D., Hybrid Monte Carlo. *Phys. Lett. B* **195**, 216-222 (1987)