Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky
Katedra informatiky

Decoding of Hidden Markov Models with Applications to Sequence Alignment

Michal Nánási

---

na získanie akademického titulu *philosophiae doctor*
v odbore doktorandského štúdia: 9.2.1 Informatika

Bratislava, 2014

Dizertačná práca bola vypracovaná v dennej forme doktorandského štúdia na Katedre informatiky Fakulty matematiky, fyziky a informatiky Univerzity Komenského v Bratislave.

Predkladateľ: Michal Nánási
Katedra informatiky
Fakulta matematiky, fyziky a informatiky
Univerzita Komenského
Mlynská dolina
842 48 Bratislava

Školiteľ: doc. Mgr. Bronislava Brejová, PhD.
Katedra informatiky
FMFI UK, Bratislava

Oponenti: ....
....
....
....

....
....
....
....

....
....
....

Obhajoba dizertačnej práce sa koná dňa .... o .... hod.
pred komisiou pre obhajobu dizertačnej práce v odbore doktorandského štúdia vymenovanou predsedom odborovej komisie dňa ....

v študijnom odbore 9.2.1 Informatika

na Fakulte matematiky, fyziky a informatiky Univerzity Komenského v Bratislave,
Mlynská dolina, 842 48 Bratislava.

Predseda odborovej komisie: prof. RNDr. Branislav Rovan, PhD.
Katedra informatiky
FMFI UK, Bratislava

Self Report (Autoreferát)

Introduction

New sequencing technologies are producing more and more biological data, including genomic sequences of many species. Therefore it is important to develop tools for automated analysis of such data. In dissertation we focus on computational methods for sequence annotation and sequence alignment.

In the sequence annotation problem, we want to label parts of the sequences according to their function, or meaning. We call such a labeling an *annotation*. For example, we can label each symbol of a genomic sequence base on whether it is part of a gene or not as in the following example (g is a label representing genes and n is a label for non-gene parts).

```
Sequence:     ACGGTGCGTTAGCTGCTCTGATGTCTTCGATCTAGCTAGT
Annotation:   nnnnnnnngggggggggggggggggggnnnnnnnnnnngg
```

The *sequence alignment* is a data structure that characterizes similarity or shared origin of two or more sequences. We insert gap symbols (dashes) so that corresponding parts of the sequence are in the same column as in the following example.

```
Sequence X:   CTGCTAGCTACGT--GTGT
Sequence Y:   ---------ACGTGGAT--
```

Both annotation and alignment are fundamental bioinformatics problems. The first stages of analysis of newly sequenced genomes typically include aligning new sequences with the genomes of related species, that are already sequenced, and searching for known structures (like genes) inside new genomes. The search for know structures is done using sequence annotation methods. Many subsequent methods for analysing genomes rely on sequence annotation and sequence alignment. To avoid artefacts in the results of these downstream methods, there is need to develop algorithms for producing sequence annotation and alignment with as low error rate as possible. Tools for both sequence annotation and alignment are often based on hidden Markov models (HMMs). We propose new techniques for use of HMMs in these domains and also give proofs of NP-hardness for several related problems.

We work with generative probabilistic models, hidden Markov Models (HMM) and their variants. In general, an HMM is a state machine that generates a sequence (string) along with a sequence of states (called state path). Since an HMM is a probabilistic model, it also defines the probability of sequences and state paths. The state path contains information about the structure of the generated sequence. In practice we are often given the generated sequence and the state path is hidden. The goal of the *decoding algorithm* is to reverse the generation process and obtain the original state path or at least its approximation.

While the sequence annotation and sequence alignments seem to be very different problems, the unifying element in dissertation is the use of HMMs, and in particular the selection of decoding algorithms. Selection of appropriate decoding algorithm is often neglected in practice, and usually most of the development is focused on the structure of an HMM. However, the right selection of a decoding algorithm can lead to significant improvements int the model prediction. We summarize the main contributions in two following sections.

Sequence Annotation

We study a special type of decoding algorithms: two-stage algorithms. In the first stage, the algorithm infers important aspects of the annotation, and in the second stage it fills remaining details in a way consistent with the first-stage results. We test this approach on the HIV recombination detection problem. We extend the Viterbi algorithm and the HERD algorithm to two-stage algorithms, and we show that two-stage algorithms can improve the accuracy of decoding (as far we know, such algorithms were previously used only for reducing the running time). Then we study the computational complexity of several decoding criteria appropriate for the first stage of two-stage algorithm, and we show NP-hardness results for obtaining the optimal annotations using these criteria.

We study computational complexity problems related to footprints and sets of a state path. Footprint of state path π, denoted as $f(\pi)$, is maximal subsequence of π such that it does not contain two same consecutive states. Set of a state path, denoted as $S(\pi)$ is the set of states that are in the state path π. For example, if $\pi = uvvvuvvvuuuuvv$, then $f(\pi) = uvuvuv$ and $S(\pi) = \{u, v\}$. The probability of set/footprint K given sequence X is the sum of the probabilities of state paths generating X that have set/footprint K. We studied three following problems.

Definition 1 (The most probable set problem)**.** *Given an HMM H, sequence X of length n and a number $p \in [0, 1]$, decide if there exists a set of states S such that* $\Pr(S(\pi) = S, X \mid H) \geq p$.

Definition 2 (The most probable restriction problem)**.** *Given an HMM H, sequence X, integer l and number $p \in [0, 1]$, decide if there exists a subset of states S of size l, such that* $\Pr(S(\pi) \subseteq S, X \mid H) \geq p$.

Definition 3 (The most probable footprint problem)**.** *Given an HMM H, sequence X of length n and a number $p \in [0, 1]$, decide if there exists a footprint F such that* $\Pr(f(\pi) = F, X \mid H) \geq p$.

We showed that the most probable footprint problem and the most probable restriction problems are NP-complete and that the most probable set problem is NP-hard. It is not clear if the most probable

Algorithm	Alignment error	Repeat sn.	Repeat sp.	Block sn.	Block sp.
3-state VA (baseline)	4.78%				
3-state PD	**4.41%**				
Context	5.98%				
Muscle	5.62%				
SFF MPD	**3.37%**	**95.97%**	97.78%	**43.07%**	44.87%
SFF PD	3.53%	95.86%	97.87%	42.70%	47.37%
SFF BPD	3.51%	93.09%	**98.07%**	36.50%	41.67%
SFF BVA	3.91%	93.26%	97.96%	35.77%	40.66%
SFF VA	4.04%	95.29%	97.85%	42.70%	**48.95%**

Tabuľka 1: Accuracy of decoding methods on simulated data. Best result in each metric is bold. Upper part of the table contain standard 3-state HMM for aligning sequences with the Viterbi algorithm and Posterior decoding. Additionally we also include alignment software Muscle and Context. Standard algorithms methods do not provide repeat annotation and therefore only error rate is available for them.

set problem is in NP, because even deciding if the probability of a set is non-zero is NP-hard. We show different proofs for these theorems, and then we show the modification of proof of the most probable footprint problem to proofs of the other two theorems.

Sequence Alignment

In sequence alignment, the goal is to search for corresponding parts of the sequences and arrange them into same position in the alignment. To choose the biologically correct alignment, we usually optimize some scoring scheme. We will consider scoring schemes which are defined using pair hidden Markov models (pHMM). A pHMM generates pairs of sequences along with their alignment (an alignment is defined by the state path). This model is an extension of HMM.

We proposed a tractable method for aligning sequences with tandem repeats. A tandem repeat consists of consecutive copies (not exact) of a certain motif (short genomic sequence). Tandem repeats cause problems with sequence alignments because it is hard to distinguish between individual copies of the motif. We extend a traditional pHMM model for sequence alignment by additional states modeling tandem repeats. We call this model Sunflower field model (SFF). We propose new decoding algorithms tailored to this model, namely block Viterbi algorithm (BVA), and block posterior decoding (BPD). The method was tested on simulated data that resembles human-dog alignemnt. We run SFF with standard decoding algorithms, the Viterbi algorithm (VA), posterior decoding (PD), marginalized posterior decoding (MPD), and with the two new decoding algorithms. One such comparison is in table 1, where we compare algorithms using different accuracy measures. The alignment error is the fraction of incorrectly predicted columns of an alignment. Repeat sensitivity and specificity measures the accuracy of prediction of repeat annotation for individual bases. The block sensitivity and specificity measures the accuracy of finding the tandem repeats with exact boundaries. The sensitivity is the number of correctly predicted features divided by the number of correct features. The specificity is the number of correctly predicted features divided by the number of all predicted features. The results showed that our new model and decoding methods decreased the error rate. In particular, method increased accuracy near the border of tandem repeats (such data are not shown here).

Paper about this method was presented on *WABI 2013* conference, journal version was published in *Algorithms in Molecular Biology*. Journal version contain also experimental evaluation on real sequences.

Bibliography

[Aggarwal et al., 1987] Aggarwal, A., Klawe, M., Moran, S., Shor, P., and Wilber, R. (1987). Geometric Applications of a Matrix-Searching Algorithm. *Algorithmica*, 2:195–208.

[Alexandersson et al., 2003] Alexandersson, M., Cawley, S., and Pachter, L. (2003). SLAM: Cross-Species Gene Finding and Alignment with a Generalized Pair Hidden Markov Model. *Genome Res.*, 13:496–502.

[Altschul et al., 1990] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic Local Alignment Search Tool. *J Mol Biol*, 215(3):403–10.

[Altschul et al., 1997] Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a New Generation of Protein Database Search Programs. *Nucleic Acids Res*, 25(17):3389–402.

[Arlazarov et al., 1970] Arlazarov, V. L., Dinic, E. A., Kronrod, M. A., and Faradžev, I. A. (1970). On economical construction of the transitive closure of a directed graph. *Soviet Mathematics—Doklady*, 11(5):1209–1210.

[Benson, 1997] Benson, G. (1997). Sequence alignment with tandem duplication. *Journal of Computational Biology*, 4(3):351–357.

[Benson, 1999] Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research*, 27(2):573–580.

[Bérard et al., 2006] Bérard, S., Nicolas, F., Buard, J., Gascuel, O., and Rivals, E. (2006). A fast and specific alignment method for minisatellite maps. *Evolutionary Bioinformatics Online*, 2:303.

[Birney et al., 2004] Birney, E., Clamp, M., and Durbin, R. (2004). GeneWise and Genomewise. *Genome Res.*, 14:988–995.

[Bradley et al., 2009] Bradley, R. K., Roberts, A., Smoot, M., Juvekar, S., Do, J., Dewey, C., Holmes, I., and Pachter, L. (2009). Fast Statistical Alignment. *PLoS Comput Biol*, 5(5):e1000392.

[Bray et al., 2003] Bray, N., Dubchak, I., and Pachter, L. (2003). AVID: A Global Alignment Program. *Genome Res*, 13(1):97–102.

[Brazma et al., 2001] Brazma, A., Parkinson, H., Schlitt, T., and Shojatalab, M. (2001). A quick introduction to elements of biology - cells, molecules, genes, functional genomics, microarrays. http://www.ebi.ac.uk/microarray/biology_intro.html.

[Brejová et al., 2005] Brejová, B., Brown, D. G., Li, M., and Vinař, T. (2005). ExonHunter: a comprehensive approach to gene finding. *Bioinformatics*, 21 Suppl 1:i57–65.

[Brejová et al., 2005] Brejová, B., Brown, D. G., and Vinař, T. (2005). Vector seeds: An extension to spaced seeds. *Journal of Computer and System Sciences*, 70(3):364 – 380.

[Brejová et al., 2007] Brejová, B., Brown, D. G., and Vinař, T. (2007). The most probable annotation problem in HMMs and its application to bioinformatics. *J. Comput. Syst. Sci.*, 73:1060–1077.

[Brown and Truszkowski, 2010] Brown, D. G. and Truszkowski, J. (2010). New decoding algorithms for hidden Markov models using distance measures on labellings. In *Asia Pacific Bioinformatics Conference (APBC)*.

[Burge and Karlin, 1997] Burge, C. and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology*, 268(1):78 – 94.

[Cartwright, 2009] Cartwright, R. A. (2009). Problems and solutions for estimating indel rates and length distributions. *Mol Biol Evol*, 26(2):473–80.

[Chao et al., 1992] Chao, K. M., Pearson, W. R., and Miller, W. (1992). Aligning two sequences within a specified diagonal band. *Comput Appl Biosci*, 8(5):481–7.

[Crochemore et al., 2002] Crochemore, M., Landau, G. M., and Ziv-Ukelson, M. (2002). A sub-quadratic sequence alignment algorithm for unrestricted cost matrices. In *Proceedings of the thirteenth annual ACM-SIAM symposium on Discrete algorithms*, SODA '02, pages 679–688, Philadelphia, PA, USA. Society for Industrial and Applied Mathematics.

[Csűrös and Ma, 2005] Csűrös, M. and Ma, B. (2005). Rapid Homology Search with Two-Stage Extension and Daughter Seeds. In Wang, L., editor, *Computing and Combinatorics*, volume 3595 of *Lecture Notes in Computer Science*, pages 104–114. Springer Berlin / Heidelberg.

[Durbin et al., 1998] Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. (1998). *Biological sequence analysis: Probabilistic models of proteins and nucleic acids*. Cambridge University Press.

[Edgar, 2004] Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5):1792–1797.

[Freschi and Bogliolo, 2012] Freschi, V. and Bogliolo, A. (2012). A lossy compression technique enabling duplication-aware sequence alignment. *Evolutionary Bioinformatics Online*, 8:171.

[Frith, 2011] Frith, M. C. (2011). A new repeat-masking method enables specific detection of homologous sequences. *Nucleic Acids Res*, 39(4):e23.

[Garey and Johnson, 1990] Garey, M. R. and Johnson, D. S. (1990). *Computers and Intractability; A Guide to the Theory of NP-Completeness*. W. H. Freeman & Co., New York, NY, USA.

[Gemayel et al., 2010] Gemayel, R., Vinces, M. D., Legendre, M., and Verstrepen, K. J. (2010). Variable tandem repeats accelerate evolution of coding and regulatory sequences. *Annual Review of Genetics*, 44:445–477.

[Gill and Nyu, 2006] Gill, O. and Nyu, C. I. (2006). PLANAR: RNA Sequence Alignment using Non-Affine Gap Penalty and Secondary Structure.

[Gill et al., 2004] Gill, O., Zhou, Y., and Mishra, B. (2004). Aligning Sequences with Non-Affine Gap Penalty: PLAINS Algorithm, a Practical Implementation, and its Biological Applications in Comparative Genomics. In *Advances In Bioinformatics And Its Applications*.

[Grice et al., 1997] Grice, J. A., Hughey, R., and Speck, D. (1997). Reduced space sequence alignment. *Comput Appl Biosci*, 13(1):45–53.

[Gross et al., 2007] Gross, S. S., Do, C. B., Sirota, M., and Batzoglou, S. (2007). CONTRAST: a discriminative, phylogeny-free approach to multiple informant de novo gene prediction. *Genome Biology*, 8(12):R269.

[Guindon and Gascuel, 2003] Guindon, S. and Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*, 52(5):696 – 704.

[Gusfield, 2007] Gusfield, D. (2007). *Algorithms on Strings, Trees, and Sequences : Computer Science and Computational Biology*. Cambridge Univ. Press.

[Hickey and Blanchette, 2011] Hickey, G. and Blanchette, M. (2011). A probabilistic model for sequence alignment with context-sensitive indels. *Journal of Computational Biology*, 18(11):1449–1464.

[Hirschberg, 1975] Hirschberg, D. (1975). A linear space algorithm for computing maximal common subsequences. *Commun. ACM*, 18:341–343.

[Hudek and Brown, 2011] Hudek, A. and Brown, D. (2011). FEAST: Sensitive Local Alignment with Multiple Rates of Evolution. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, 8(3):698 –709.

[Hudek, 2010] Hudek, A. K. (2010). *Improvements in the Accuracy of Pairwise Genomic Alignment*. PhD thesis, University of Waterloo.

[Kall et al., 2005] Kall, L., Krogh, A., and Sonnhammer, E. L. L. (2005). An HMM posterior decoder for sequence feature prediction that includes homology information. *Bioinformatics*, 21 Suppl 1:i251–257.

[Keibler et al., 2007] Keibler, E., Arumugam, M., and Brent, M. R. (2007). The Treeterbi and Parallel Treeterbi algorithms: efficient, optimal decoding for ordinary, generalized and pair HMMs. *Bioinformatics*, 23(5):545–54.

[Kent, 2002] Kent, W. J. (2002). BLAT–the BLAST-like alignment tool. *Genome Res*, 12(4):656–64.

[Kolpakov et al., 2003] Kolpakov, R., Bana, G., and Kucherov, G. (2003). mreps: efficient and flexible detection of tandem repeats in DNA. *Nucleic Acids Research*, 31(13):3672–3678.

[Kováč et al., 2012] Kováč, P., Brejová, B., and Vinař, T. (2012). Aligning sequences with repetitive motifs. In *Information Technologies - Applications and Theory (ITAT)*, pages 41–48.

[Krogh et al., 2001] Krogh, A., Larsson, B., Heijne, G. v., and Sonnhammer, E. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *Journal of Molecular Biology*, 305(3):567 – 580.

[Lember and Koloydenko, 2010] Lember, J. and Koloydenko, A. (2010). A constructive proof of the existence of viterbi processes. *Information Theory, IEEE Transactions on*, 56(4):2017 –2033.

[Lempel and Ziv, 1976] Lempel, A. and Ziv, J. (1976). On the Complexity of Finite Sequences. *Information Theory, IEEE Transactions on*, 22(1):75 – 81.

[Levin et al., 2006] Levin, D. A., Peres, Y., and Wilmer, E. L. (2006). *Markov chains and mixing times*. American Mathematical Society.

[Lior et al., 2004] Lior, P., Marina, A., and Cawley, S. (2004). Applications of Generalized Pair Hidden Markov Models to Alignment and Gene Finding Problems. *Journal of Computational Biology*, 9.

[Liu et al., 2010] Liu, Y., Schmidt, B., and Maskell, D. L. (2010). MSAProbs: multiple sequence alignment based on pair hidden Markov models and partition function posterior probabilities. *Bioinformatics*, 26(16):1958–64.

[Lu et al., 2009] Lu, D., Brown, R., Arumugam, M., and Brent, M. (2009). Pairagon: a highly accurate, HMM-based cDNA-to-genome aligner. *Bioinformatics*, 25:1587–1593.

[Lunter et al., 2008] Lunter, G., Rocco, A., Mimouni, N., Heger, A., Caldeira, A., and Hein, J. (2008). Uncertainty in homology inferences: assessing and improving genomic sequence alignment. *Genome Res*, 18(2):298–309.

[Lyngsø and Pedersen, 2002] Lyngsø, R. B. and Pedersen, C. N. S. (2002). The consensus string problem and the complexity of comparing hidden Markov models. *Journal of Computer and System Sciences*, 65(3):545 – 569.

[Ma et al., 2002] Ma, B., Tromp, J., and Li, M. (2002). PatternHunter: faster and more sensitive homology search. *Bioinformatics*, 18(3):440–5.

[Majoros et al., 2005] Majoros, W., Pertea, M., and Salzberg, S. (2005). Efficient implementation of a generalized pair hidden Markov model for comparative gene finding. *Bioinformatics*, 21:1782–1788.

[Majoros et al., 2004] Majoros, W. H., Pertea, M., and Salzberg, S. L. (2004). TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics*, 20(16):2878–9.

[Messer and Arndt, 2007] Messer, P. W. and Arndt, P. F. (2007). The majority of recent short DNA insertions in the human genome are tandem duplications. *Mol Biol Evol*, 24(5):1190–1197.

[Meyer and Durbin, 2002] Meyer, I. and Durbin, R. (2002). Comparative ab initio prediction of gene structures using pair HMMs. *Bioinformatics*, 18:1309–1318.

[Myers and Miller, 1989] Myers, E. and Miller, W. (1989). Approximate Matching of Regular Expressions. *Bulletin of Mathematical Biology*, 51(1):5–37.

[Nánási et al., 2010] Nánási, M., Vinař, T., and Brejová, B. (2010). The Highest Expected Reward Decoding for HMMs with Application to Recombination Detection. In Amir, A. and Parida, L., editors, *Combinatorial Pattern Matching, 21th Annual Symposium (CPM 2010)*, volume 6129 of *Lecture Notes in Computer Science*, pages 164–176, Brooklyn, New York, USA. Springer.

[Nánási et al., 2014] Nánási, M., Vinař, T., and Brejová, B. (2014). Probabilistic approaches to alignment with tandem repeats. *Algorithms for Molecular Biology*, 9(1):3.

[Nánási, 2010] Nánási, M. (2010). Biological sequence annotation with hidden Markov models. Master's thesis, Comenius University.

[Pachter et al., 2002] Pachter, L., Alexandersson, M., and Cawley, S. (2002). Applications of generalized pair hidden Markov models to alignment and gene finding problems. *Journal of Computational Biology*, 9(2):389–399.

[Robertson et al., 2000] Robertson, D. L., Anderson, J. P., Bradac, J. A., Carr, J. K., Foley, B., Funkhouser, R. K., Gao, F., Hahn, B. H., Kalish, M. L., Kuiken, C., Learn, G. H., Leitner, T., McCutchan, F., Osmanov, S., Peeters, M., Pieniazek, D., Salminen, M., Sharp, P. M., Wolinsky, S., and Korber, B. (2000). HIV-1 nomenclature proposal. *Science*, 288(5463):55–6.

[Sammeth and Stoye, 2006] Sammeth, M. and Stoye, J. (2006). Comparing tandem repeats with duplications and excisions of variable degree. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 3(4):395–407.

[Satija et al., 2010] Satija, R., Hein, J., and Lunter, G. A. (2010). Genome-wide functional element detection using pairwise statistical alignment outperforms multiple genome footprinting techniques. *Bioinformatics*, 26(17):2116–20.

[Schultz et al., 2006] Schultz, A.-K., Zhang, M., Leitner, T., Kuiken, C., Korber, B., Morgenstern, B., and Stanke, M. (2006). A jumping profile Hidden Markov Model and applications to recombination sites in HIV and HCV genomes. *BMC Bioinformatics*, 7:265.

[Šrámek et al., 2007] Šrámek, R., Brejová, B., and Vinař, T. (2007). On-Line Viterbi Algorithm for Analysis of Long Biological Sequences. In Giancarlo, R. and Hannenhalli, S., editors, *Algorithms in Bioinformatics*, volume 4645 of *Lecture Notes in Computer Science*, pages 240–251. Springer Berlin / Heidelberg.

[Truszkowski and Brown, 2011] Truszkowski, J. and Brown, D. G. (2011). More accurate recombination prediction in HIV-1 using a robust decoding algorithm for HMMs. *BMC Bioinformatics*, 12:168.

[Vinař, 2005] Vinař, T. (2005). *Enhancements to Hidden Markov Models for Gene Finding and Other Biological Applications*. PhD thesis, University of Waterloo.

[Weimann, 2009] Weimann, O. (2009). *Accelerating dynamic programming*. PhD thesis, Massachusetts Institute Of Technology, Cambridge, MA, USA.

[Wexler et al., 2005] Wexler, Y., Yakhini, Z., Kashi, Y., and Geiger, D. (2005). Finding approximate tandem repeats in genomic sequences. *Journal of Computational Biology*, 12(7):928–942.

[Zvelebil and Baum, 2007] Zvelebil, M. and Baum, J. (2007). *Understanding Bioinformatics*. Garland Science, 1 edition.

Abstract

We study two important problems in computational biology: sequence annotation and sequence alignment. In the thesis we concentrate on the use of hidden Markov models (HMMs), well established generative probabilistic models.

In the first part, we study the sequence annotation problem, specifically the two-stage HMM decoding algorithms and the computational complexity of related problems. In particular, we demonstrate that two-stage algorithms can be used to increase the accuracy of decoding, and we prove the NP-hardness for three problems appropriate for the first stage: the most probable set problem, the most probable restriction problem and the most probable footprint problem.

The second part of the thesis focuses on alignment of sequences that contain tandem repeats. Tandem repeats are highly repetitive elements withing genomic sequences that cause biases in alignments. To address this issue, we introduce a new HMM that models alignments containing tandem repeats, combine it with existing and new decoding algorithms, and evaluate our approach experimentally.

In both problems, we use the decoding algorithms to improve the accuracy of HMM predictions. Decoding algorithms are often neglected, and most of the development is focused on the structure of an HMM. However, a proper selection of a decoding method can lead to significant improvements in the predictions.

Abstrakt

Zaoberáme sa dvomi dôležitými bioinformatickými problémami: anotáciou sekvencií a zarovnávaním sekvencií. V práci sa sústredíme na využitie skrytých Markovových modelov (HMM), dobre známych generatívnych pravdepodobnostných modelov.

V prvej časti študujeme anotáciu sekvencií, konkrétne dvojstupňové dekódovacie algoritmy a výpočtové problémy, ktoré s nimi súvisia. Ukážeme, že dvojstupňové algoritmy môžu zlepšiť presnosť dekódovania a dokážeme, že tri problémy vhodné pre prvý stupeň výpočtu sú NP-ťažké: problém najpravdepodobnejšej množiny, problém najpravdepodobnejšej reštrikcie a problém najpravdepodobnejšej stopy.

Druhá časť sa zaoberá zarovnávaním sekvencií, ktoré obsahujú tandemové opakovania. Tandemové opakovania sú opakujúce sa časti genomických sekvencií, ktoré často spôsobujú chyby v zarovnaniach. Aby sme vyriešili tento problém, vyvinuli sme nový HMM, ktorý modeluje zarovnania obsahujúce tandemové opakovania a skombinovali sme ho s existujúcimi ako aj novými dekódovacími algoritmami. Náš prístup sme vyhodnotili experimentálne.

V oboch problémoch sme používali dekódovacie algoritmy na zlepšenie presnosti predikcií HMM. Dekódovacie algoritmy sú často podceňované a väčšina vývoja ide do vytvárania topológie HMM. Avšak správnym výberom dekódovacej metódy môžeme dosiahnuť významné zlepšenie predikcií.

Vlastné publikácie autora

- Nánási, M., Vinař, T., Brejová, B.,(2014) Probabilistic approaches to alignment with tandem repeat. In *Algorithms for Molecular Biology.* Vol. 9, No. 1 (2014), Art. No. 3, s. 1-11, http://www.almob.org/content/9/1/3

  Conference version published in *Algorithms in Bioinformatics*, WABI 2013, Volume 8126 of Lecture Notes in Computer Science, Springer, 2013. pages 287-299,

- Nánási, M., Vinař, T., Brejová, B.,(2012) Sequence Annotation with HMMs: New Problems and Their Complexity. *Preprint*, http://arxiv.org/abs/1210.2587

- Nánási, M., Vinař, T., Brejová, B., (2010) The highest expected reward decoding for HMMs with application to recombination detection In: Combinatorial Pattern Matching, Volume 6129 of Lecture Notes in Computer Science, pages 164-176, Springer-Verlag

  - Truszkowski, J. - Brown, D. G. (2011). More accurate recombination prediction in HIV-1 using a robust decoding algorithm for HMMs. In: BMC Bioinformatics, Vol. 12, 2011, Art. No. 168
  - Hamada, M. - Asai, K. (2012). A classification of bioinformatics algorithms from the viewpoint of maximizing expected accuracy (MEA). In: Journal of Computational Biology, Vol. 19, No. 5, 2012, s. 532-549