IRODALOMJEGYZÉK

- Adachi, J. & Hasegawa, M. (1990) Amino acid substitution of proteins coded for in mithocondrial DNA during mammalian evolution. *Jpn. J. Genet.* **67:**187-197.
- Anfinsen, C.D. (1973) Principles that govern the folding of protein chains. *Scianece*, **181:**223-230.
- Barnhart, M.M., Pinker, J.S., Soto, G.E., Sauer, F.G., Langerman, S., Waksman, G. Frieden, C. & Hultgren, S.J. (2000) PapD-like chaperones provide the missing information for folding proteins. *Proc. Natl. Academy Science* **97:**7709-7714.
- Bellman, R. (1957) Dynamic programming. University Press, Princetown.
- Benner, S.A., Cohen, M.A. & Gonnet, G.H. (1993) Empirical and structural models for insertions and deletions in the divergent evolution of proteins. *J. Mol. Biol.* **229:**1065-1082.
- Bishop, M. J. & Thompson, E. A. (1986) Maximum likelihood alignment of DNA sequences. *J. Mol. Biol.* **190:**159-165.
- Cao, Y., Adachi, J., Janke, A., Paabo, S. & Hasegawa, M. (1994) Phyilogenetic relationships among eutherian orders estimated from inferred sequences of mithocondrial proteins: instability of a tree based on a single gene. *J. Mol. Evol.* **39:**519-527.
- Carillo, H. & Lipman, D. (1988) The multiple sequence alignment problem in biology. *SIAM J. Appl. Math.* **48:**1073-1082.
- Cavalli-Sforza, L.L. & Edwards, A.W.F. (1967) Phylogenetic analysis: models and estimation procedures. *Evolution* **32:**550-570.
- Corpet, F. (1988) Multiple Sequence alignment with hierarchical clustering. *Nucleic Acids Res.* **16:**10881-10890.
- Cox, D.R. & Smith, W.L. (1961) Queues. McGraw-Hill, New York.
- Day, H.E.W (1983) Computationally difficult parsimony problems in phylogenetic systematics *J. theor. Biol.* **103:**429-438.
- Dayhoff, M.O., Schwartz, R.M. & Orcutt, B.C. (1978) A model of evolutionary change in proteins. *Atles of Protein Sequence and Structure* **5:**345-352.
- Durbin, R., Eddy, S., Krogh, A. & Mitchinson, G. (1998) Biological sequence analysis. University Press, Cambridge.
- Feller, W. (1968) An introduction to the probability theory and its applications, Vol. 1. McGraw-Hill, New York.
- Felsenstein, J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* **17:**368-376.
- Feng, D. & Doolittle, R.F. (1987) Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J. Mol. Evol.* **25:**351-360.
- Fickett, J.W. (1984) Fast optimal alignment. *Nucleic Acids Res.* **12:**175-180.
- Fleiβner, R., Metzler, D. & von Haeseler, A. (2000) Can one estimate distances from pairwise sequence alignments? In: Bornberg-Bauer, E., Rost, U., Stoye, J., Vingron, M. (eds) GCB 2000, Proceedings of the German Conference on Bioinformatics, Oct 5-7. 2000, Heidelberg. Logos Verlag, Berlin, 89-95.
- Foulds, L.R. & Graham, R.L. (1982) The Steiner problem in phylogeny is NP-complete. *Advances Appl. Math.*, **3:**43-49.

- Galil, Z. & Giancarlo, R. (1989) Speeding up dynamic programming with applications to molecular biology. *Theor. Comp. Sci.* **64:**107-118.
- Gamerman, D. (1997) Markov Chain Monte Carlo. Chapman & Hall. London.
- Gonnet, G.H., Cohen, M.A. & Benner, S.A. (1992) Exhaustive matching of the entire protein sequence database. *Science* **256:**1443-1445.
- Gotoh, O. (1982) An improved algorithm for matching biological sequences. *J. Mol. Biol.* **162:**705-708.
- Gribiskov, M., McLachan, A. & Eisenberg, D. (1987) Profile analysis detection of distantly related proteins. *Proc. Natl. Academy Science* **88:**4355-4358.
- Gusfield, D. (1993) Efficient methods for multiple sequence alignment with guaranteed error bounds. *Bull. Math. Biol.* **55:**141-154.
- Gusfield, D (1997) Algorithms on strings, trees and sequences. University Press, Cambridge.
- Hartigan, J.(1975) Clustering Algorithms. Wiley, New York.
- Hasegawa, M. & Fujiwara, M. (1993) Relative efficiencies of the maximum likelihood, maximum parsimony, and neighbor-joining methods for estimating protein phylogeny. *Mol. Phylogenet. Evol.* **2:**1-5.
- Hasegawa, M., Kishino, H. & Yano, T. (1985) Dating the human-ape splitting by a molecular clock of mitochondrial RNA. *J. Mol. Evol.* **22:**160-174.
- Hein, J. (2001) An algorithm for statistical alignment of sequences related by a binary tree. In: *Pacific Symposium on Biocomputing,* (R.B.Altman, A.K.Dunker, L.Hunter, K.Lauderdale & T.E.Klein, eds.) 179-190. Singapore: World Scientific.
- Hein, J., Wiuf, C., Knudsen, B., Moller, M.B., Wiblig, G. (2000) Statistical alignment: computational properties, homology testing and goodness-of-fit. *J. Mol. Biol.* **302:**265-279.
- Holmes, I. & Bruno, W.J.: (2001) Evolutionary HMMs: a Bayesian approach to multiple alignment. *Bioinformatics*, **accepted**.
- Hubbard, T.J.P., Lesk, A.M. & Tramontano, A. (1996) Gathering them into the fold. *Nature Structural Biology* **4:**313.
- Jones, D.T., Taylor, W.R. & Thornton, J.M. (1992) The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.* **8:**275-282.
- Jukes, T.H. & Cantor, C.R. (1969) Evolution in protein molecules. In: *Mammalian Protein Metabolism.* (H.N.Munro, ed.) 21-123. Academic, New-York.
- Kimura, M. (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16:**11-120.
- Kimura, M. (1981) Estimation of evolutionary distances between homologous nucleotide sequences. *Proc. Natl. Academy Science* **78:**454-458.
- Kishino, H. & Hasegawa, M. (1990) Converting distance to time: application to human evolution. *Methods Enzymol.* **183:**550-570.
- Kishino, H., Miyata, T. & Hasegawa, M. (1990) Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. *J. Mol. Evol.* **31:**151-160.
- Lipman, D., Altshul, S. & Kececioglu, J. (1989) A tool for multiple sequence alignment. *Proc. Natl. Academy Science* **86:**4412-4415.
- Metzler, D., Fleiβner, R., von Haeseler, A. & Wakolbinger, A. (2001) Assessing variability by joint sampling of alignments and mutation rates. *J. Mol. Evol.* **to appear.** http://www.math.uni-frankfurt.de/~stoch/software/mcmcalgn/

- Miklós, I. (2001a) Irreversible likelihood models. *European Mathematical Genetics Meeting, 2001. Apr. 20-21, Lille, France.*
- Miklós, I. (2001b) An improved model for statistical alignment of sequences evolved by a star tree. *Phylogenetics Combinatorics, 2001. Jun. 10-16, Bielefeld, Germany.*
- Miklós, I. (2001c) Algorithm for statistical alignment of sequences derived from a Poisson sequence length distribution. *Disc. Appl. Math.* **accepted.**
- Miklós, I. & Toroczkai, Z. (2001) An improved model for statistical alignment, in: *WABI2001, Lecture Notes in Computer Science,* (O.Gascuel & B.M.E.Moret, eds.) **2149:**1-10. Springer, Berlin.
- Miller, W. & Myers, E.W. (1988) Sequence comparison with concave weighting functions. *Bull. Math. Biol.* **50:**97-120.
- Morgenstern, B, Dress, A. & Werner, T. (1996) Multiple DNA and protein sequence alignment based on segment-to-segment comparison. *Proc. Natl. Academy Science* **93:**12098-12103.
- Morgenstern, B., Frech, K., Dress, A. & Werner, T. (1998) DIALIGN: Finding local similarities by multiple sequence alignment. *Bioinformatics* **14:**290-294.
- Morgenstern, B. (1999) DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics* **15:**211-218.
- Myers,E.W. (1986) An O(ND) difference algorithm and its variations. *Algorithmica,* **1:**251-266.
- Needleman, S.B. & Wunsch, C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48:**443-453.
- Normark, S. (2000) Anfinsen comes out of the cage during assembly of bacterial pilus. *Proc. Natl. Academy Science* **97:**7670-7672.
- Obádovics, J.Gy. & Szarka, Z. (1999) Felsőbb matematika. Scolar, Budapest.
- Pages & Holmes (1998) Molecular Evolution Phylogenetic Approach. Lackwell, Oxford.
- Podani, J. (1997) Bevezetés a többváltozós biológiai adatfeltárás rejtelmeibe. Scientia, Budapest.
- Ravi, R. & Kececioglu, J. D. (1998) Approximation algorithms for multiple sequence alignment under a fixed evolutionary tree. *Disc. Appl. Math.* **88:**355-366.
- Rodrígez, F., Oliver, J.L., Marin, A. & Medina, J.R. (1990) The general stochastic model of nucleotide substitution *J. theor. Biol.* **142:**485-501.
- Sankoff, D. (1975) Minimal mutation trees of sequences. *SIAM J. Apl. Math.* **28:**35-42.
- Sankoff, D., Cedergren, R.J. & Lapalme, G. (1976) Frequency of insertion-deletion, and transition in the evolution of 5S ribosomal RNA. *J. Mol. Evol.* **48:**443-453.
- Schwartz, R. & Dayhoff, M. (1979) Matrices for detecting distant relationships. In *Atlas of Protein Sequences* pp 353-358. Natl. Biomed. Res. Found.
- Sellers, P.H. (1974) On the theory and computation of evolutionary distances. *SIAM J. Appl. Math.* **26:**787-793.
- Smith, T.F. & Waterman, M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.* **147:**195-197.
- Smith, T.F., Waterman, M.S. & Fitch, W.M. (1981) Comparative biosequence metrics. *J. Mol. Evol.* **18:**38-46.
- Sneddon, I.N. (1957) Elements of partial differential equations. McGraw-Hill, New-York.
- Spouge, J.L. (1989) Speeding up dynamic programming algorithms for finding optimal lattice paths. *SIAM J. Appl. Math.* **49:**1552-1566.
- Spouge, J.L. (1991) Fast optimal alignment. *CABIOS* **7:**1-7.

- Steel, M. & Hein, J. (2001) Applying the Thorne-Kishino-Felsenstein model to sequence evolution on a star-shaped tree. *Appl. Math. Lett.* **14:**679-684.
- Tamura, K. & Nei, M. (1993) Estimation of the number of nucleotide substitutions in the control region of mithocondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.* **10:**512-526.
- Thompson, J.D., Higgins, D.G. & Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific penalties and weight matrix choice. *Nucleic Acids Res.* **22:**4673-4680.
- Thorne, J.L., Kishino, H. & Felsenstein, J. (1991) An evolutionary model for Maximum likelihood alignment of DNA sequences. *J. Mol. Evol.* **33:**114-124.
- Thorne, J.L., Kishino, H. & Felsenstein, J. (1992) Inching toward reality: an improved likelihood model of sequence evolution. *J. Mol. Evol.* **34:**3-16.
- Ukkonen, E. (1984) On approximate string matching. In *Proc. Int. Conf. Found. Comp. Theor., Lecture Notes Comput. Sci.* **158:**487-495.
- Ukkonen, E. (1985) Algorithms for approximate string matching. *Inform. Control* **64:**100-118.
- Wang, L. & Jiang, T. (1994) On the complexity of multiple sequence alignment. *J. Comp. Biol.* **1:**337-348.
- Waterman, M.S. (1984) Efficient sequence alignment algorithms. *J. theor. Biol.* **108:**333-337.
- Waterman. M.S., Smith, T.F. & Beyer, W.A. (1976) Some biological sequence metrics. *Advan. Math.* **20:**367-387.
- Watson, J.D. & Crick, F.H.C. (1953) Genetical implications of the structure of the dezoxyribonucleic acid. *Nature* **171:**964-967.
- Wu,S., Manber,U., Myers,G. and Miller,W. (1990) An O(NP) sequence comparison algorithm *Information Processing Letters*, **35:**317-323.
- Zhang, J. (2000) Protein length distributions for the three domains of life, *Trends Genet.* **16:**107-109.
- Zharkikh, A. (1994) Estimation of evolutionary distances between nucleotide sites. *J. Mol. Evol.* **39:**315-329.

KÖSZÖNETNYILVÁNÍTÁS

Szeretnék köszönetet mondani Jeff Thorne-nak, Carsten Wiufnak, Jotun Heinnek és Ian Holmesnak a rendkívül hasznos diszkussziókért. Az ő segítségük nagy mértékben hozzájárult ahhoz, hogy ez az értekezés itt, Magyarországon elkészülhessen. Külön köszönet Jotun Heinnek és Ian Holmesnak, amiért publikáció előtti kéziratait elküldték nekem, így lehetővé téve, hogy megismerhessem a statisztikus szekvencia illesztés legújabb eredményeit.

Köszönet témavezetőmnek, Podani Jánosnak, a hasznos kommentárokért, kritikákért, a kéziratok gondos áttanulmányozásáért és javításáért.

Szeretném megköszönni szerzőtársamnak, Toroczkai Zoltánnak, a közös munka örömét, valamint azt, hogy felhívta a figyelmemet a sztochasztikus sorban állási rendszerek és a beszúrás-törlés modellek közötti párhuzamra.

A bemutatott sarokvágási technikák algoritmusainak programozásában Kun Ádám volt segítségemre. Köszönet érte.

Köszönet illeti továbbá az alábbi személyeket, akik számos apró segítséget nyújtottak az értekezés elkészültében: Márton Zsuzsanna, Dr. Márton Józsefné dr., Pál Csaba, Scheuring István, Obornyi Beáta.

RÖVID ÖSSZEFOGLALÁS

Az elmúlt két évtizedben a szubsztitúciók modellezése látványosan fejlődött. A beszúrások és törlések modellezésére nem fordítottak nagy figyelmet, és jelenleg is hasonlóság/távolság alapú módszerekkel történik a biológiai szekvenciák illesztése. Azonban először távolság/hasonlóság alapú módszerrel illeszteni, majd az analízis felénél átváltani statisztikai módszerre inkonzisztens megközelítés. Egy rossz illesztés hibát okozhat az evolúciós paraméterek meghatározásában, másrészről pontatlan paraméterek pontatlan illesztéshez vezetnek.

A statisztikus illesztés módszere sztochasztikusan modellezi mind a szubsztitúciókat, mind a beszúrásokat és törléseket. Statisztikus illesztésre idáig a Thorne-Kishino-Felsenstein modellt használták. Azonban ennek a modellnek van két biológiailag irrealisztikus tulajdonsága. Ez a modell nem enged meg többszörös beszúrásokat és törléseket egyetlen evolúciós lépésben, valamint a szekvenciák hosszúságának az eloszlása geometriai ebben a modellben.

A disszertációban biológiailag relevánsabb modelleket mutattam be. A bemutatott modellek két nagy csoportra oszthatóak. A kombinatorikus modellek az ősi szekvenciákra tesznek különböző feltételeket, de megőrzik a TKF modell tranziens viselkedését. Bemutattam egy $O(l^3)$ futási idejű algoritmust, amely két olyan szekvencia kapcsoltsági valószínűségét számolja ki, amelyek Poisson szekvencia hosszúságból evolválódtak. Egy másik modell egy javított fragmentum modell, amely lehetővé teszi átfedő törlések modellezését. Ismertettem egy $O(l^n)$ algoritmust a többszörös statisztikus szekvencia illesztésre.

Az analitikus modellek tranziens viselkedése különbözik a TKF modellétől. Bemutattam egy olyan algoritmust, amely lehetővé teszi többszörös beszúrások modellezését, valamint egy algoritmust, amely $O(l^3)$ idő alatt kiszámolja két szekvencia kapcsoltsági valószínűségét ebben az esetben. Megmutattam, hogy hogyan lehet modellezni többszörös törléseket, valamint megadtam egy modellt, amelyben a szekvenciák hossza Poisson eloszlást követ.

Végül bemutattam, hogy hogyan lehet a sarokvágási technikát alkalmazni a statisztikus szekvencia illesztésben.

SUMMARY

During the last two decades, the analysis of the substitution process has improved considerably. The process of insertions and deletions has not received the same attention and is presently being analysed by optimisation techniques, namely, minimising distance or maximising a similarity score. However, it is an inconsistent approach to first use parsimony/distance and then halfway in the analysis switch to a statistical approach. An incorrect alignment might cause unrecognisable bias in the evolutionary parameter estimation.

The statistical alignment approach involves the stochastic modelling of the insertion-deletion and the substitution processes. The Thorne-Kishino-Felsenstein model provides a proper statistical analysis of two sequences. However, this model has two unrealistic biological properties. This model does not allow multiple insertions and deletions as a single event, and the sequence lengths have a geometric length distribution in the steady state limit.

In this dissertation, I presented several improved models for the statistical alignment, which are divided into two subsets. The models that belong to the first subset are combinatorial models. These models lay down certain conditions to the properties of ancestral sequences, but maintain the temporal behaviour of the TKF model. I introduced an $O(l^3)$ algorithm for statistical alignment of two sequences derived from a Poisson sequence length distribution, where l is the geometric average of the sequence lengths. The second model is an improved fragment model, which can model overlapping deletions. The presented algorithm computes the probability of two sequences evolved under this fragment model in $O(l^2)$ running time. Using the trick of these algorithms, I showed a faster algorithm for the multiple statistical alignment.

The models that belong to the second subset are analytical models. The temporal behaviour of these models differs from that of the TKF model. I presented a model that allows multiple insertions. The algorithm that computes the joint probability of two sequences in this case needs $O(l^3)$ running time. I gave some hints how to model multiple deletions, and showed a model that provides a peaked sequence length distribution.

Finally, I presented corner-cutting methods in the statistical alignment.