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

Algorithms for Genome Rearrangements

Jakub Kováč

---

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

Bratislava, 2013

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

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

Školiteľ: Mgr. Tomáš Vinař, PhD.
Katedra aplikovanej informatiky
FMFI UK, Bratislava

Oponenti: prof. Guillaume Fertin
LINA, Université de Nantes
2 rue de la Houssinière, BP 92208
44322 Nantes Cedex 3, Francúzsko

RNDr. Ondřej Pangrác, PhD.
Informatický ústav Univerzity Karlovy
Malostranské nám. 25
118 00 Praha 1, Česká republika

RNDr. Stefan Dobrev
Matematický ústav Slovenskej akadémie vied
Dúbravská 9, 840 00 Bratislava

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

Autoreferát

Úvod

V dizertačnej práci sa zaoberáme preusporiadaniami génov rôznych organizmov. Počas evolúcie sa genómy organizmov menia a vyvíjajú (mutujú). Okrem drobných zmien, pri ktorých sa mení len jedna alebo zopár susedných báz (nukleotidov), sa počas evolúcie z času na čas stane, že sa nejaký dlhší úsek DNA presunie na iné miesto, na opačné vlákno, či iný chromozóm. Ak sa teda dnes pozrieme na genómy príbuzných druhov, vieme v nich nájsť veľmi podobné úseky, ktoré sú však v rôznych druhoch na rôznych miestach v genóme. (Tieto rozdiely sú napríklad spoluzodpovedné za to, že sa jednotlivé druhy nemôžu krížiť.)

Vezmime si genómy dvoch druhov. V dobre definovanom matematickom modeli vieme spočítať minimálny počet mutácií (preusporiadaní), ktorý by vysvetľoval rôzne usporiadania génov v týchto genómoch. Takto dostávame akúsi mieru evolučnej príbuznosti/vzdialenosti druhov. Výhodou preusporiadaní je, že tieto „veľké“ mutácie sú v evolúcii zriedkavejšie a tak nám dovoľujú nazrieť hlbšie do histórie a porovnávať vzdialenejšie druhy, než iné tradičné prístupy, ktoré napríklad skúmajú podobnosť na úrovni nukleotidov.

Ak teda poznáme poradia génov v rôznych druhoch, môžeme si klásť nasledujúce biologicky zaujímavé otázky, ktoré sú tiež výzvou pre teoretickú informatiku:

- Nakoľko sú si dva organizmy príbuzné?
- Ako asi vyzeral genóm ich spoločného predka?
- Ak poznáme genómy viacerých druhov a ich fylogenetický strom:
  Ako asi vyzerala ich evolučná história?

Riešenia týchto problémov pomôžu biológom pri štúdiu evolučných zmien; navyše je oblasť veľmi zaujímavá aj z informatického hľadiska – obsahuje veľa algoritmických problémov, viaceré problémy sú dokázateľne ťažké a vyžadujú návrh rôznych heuristík, aproximácií, štúdium špeciálnych prípadov a podobne. Ciele tohto projektu teda sú:

- študovať teoretické problémy z oblasti preusporiadaní genómov,
- implementovať praktické riešenia a aplikovať ich na reálne dáta.

Dosiahnuté výsledky

Navrhli sme nový, biologicky hodnovernejší variant modelu DCJ. V tomto modeli riešime otázku zložitosti problémov tzv. triedenia (sorting) a pólenia genómov (genome halving). V prvom prípade sme navrhli $O(n \log n)$ algoritmus, čím sme zlepšili dovtedy známy kvadratický algoritmus, v druhom prípade sme uzavreli otvorený problém navrhnutím lineárneho algoritmu. Tieto výsledky sú spoločná práca s Robertom Warrenom, Maríliou Braga a Jensom Stoye. Boli odprezentované na konferencií RECOMB-CG 2010 v Ottawe, a v časopise *Journal of Computational Biology.*

Vyriešili sme tiež viacero otvorených teoretických problémov v breakpoint modeli: Dokázali sme, že v modeli s lineárnymi chromozómami je problém pólenia genómov NP-úplný. Navrhli sme $O(n\sqrt{n})$ algoritmus pre problém mediánu, čím sme zlepšili dovtedy známy kubický algoritmus. Tiež sme dokázali, že zlepšenie nášho algoritmu by viedlo k lepšiemu algoritmu pre hľadanie maximálneho párovania, čo je vyše 30 rokov otvorený problém. Následne sme sa venovali problému rekonštrukcie ancestrálnych usporiadaní genómov. Dokázali sme, že už pre štyri genómy je problém NP-ťažký, dokonca APX-ťažký. Poznamenajme, že sme tým vyriešili dva otvorené problémy z monografie Fertin a spol.: *Combinatorics of genome rearrangements.* Tieto výsledky boli odprezentované na konferencií RECOMB-CG 2011 v Írsku, a článok bol prijatý do časopisu *Journal of Computational Biology.*

V rámci projektu sme sa tiež venovali praktickejším otázkam v preusporiadaní genómov. Spolupracovali sme s odborníkmi z Prírodovedeckej fakulty UK, ktorí osekvenovali genómy viacerých druhov kvasiniek. Tieto genómy sú zaujímavé svojou rozmanitosťou – niektoré druhy obsahujú len jeden lineárny chromozóm, iné cirkulárny a niektoré dokonca viacero lineárnych chromozómov. Navyše niektoré genómy obsahujú duplikácie a dlhšie opakujúce sa sekvencie. Z týchto dôvodov sú tradičné existujúce riešenia ako je `MGR` (Bourque, Pevzner) a `GRAPPA` (Moret a spol.) nepoužiteľné.

Navrhli sme nový prístup k tomuto problému a implementovali sme jeden z prvých praktických nástrojov na rekonštrukciu evolučnej histórie druhov s rôznymi chromozómovými architektúrami. Vyvinutý program sme použili na rekonštrukciu usporiadaní génov kvasinkových mitochondriálnych genómov. Našu novú metódu sme porovnávali s tradičnými prístupmi. Podarilo sa nám napríklad rekonštruovať úspornejšiu evolučnú históriu chloroplastových genómov rastlín rodu *Campanulaceae* ako `MGR`, `GRAPPA`, či `ABC` (Adam, Sankoff). Tieto výsledky, spoločná práca s Tomášom Vinařom a Broňou Brejovou, boli prezentované na konferencií ISMB 2011 vo Viedni vo forme posteru a neskôr na konferencií WABI 2011 vo forme článku v recenzovanom konferenčnom zborníku. Rekonštrukciou evolučnej histórie kvasinkových genómov sme tiež prispeli do článku Matúša Valacha a kol., ktorý bol prijatý do časopisu *Nucleic Acids Research.*

Literatúra

Adam, Z. and Sankoff, D. (2008). The ABC of MGR with DCJ. *Bioinformatics*, 4:69–74.

Alekseyev, M. and Pevzner, P. (2007a). Whole genome duplications, multi-break rearrangements, and genome halving problem. In *Proc. SODA*, pages 665–679. Society for Industrial and Applied Mathematics.

Alekseyev, M. A. and Pevzner, P. A. (2007b). Colored de Bruijn graphs and the genome halving problem. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, 4(1):98–107.

Alekseyev, M. A. and Pevzner, P. A. (2007c). Whole genome duplications and contracted breakpoint graphs. *SIAM J. Comput.*, 36(6):1748–1763.

Alekseyev, M. A. and Pevzner, P. A. (2008). Multi-break rearrangements and chromosomal evolution. *Theor. Comput. Sci.*, 395(2-3):193–202.

Alimonti, P. and Kann, V. (1997). Hardness of approximating problems on cubic graphs. In *Proc. CIAC*, pages 288–298.

Atteson, K. (1997). The performance of neighbor-joining algorithms of phylogeny reconstruction. In *Proc. COCOON*, pages 101–110. Springer.

Bachrach, A., Chen, K., Harrelson, C., Mihaescu, R., Rao, S., and Shah, A. (2005). Lower bounds for maximum parsimony with gene order data. In *Proc. RECOMB-CG*, pages 1–10. Springer.

Bader, D. A., Moret, B. M. E., and Yan, M. (2001). A linear-time algorithm for computing inversion distance between signed permutations with an experimental study. *J. Comput. Biol.*, 8(5):483–491.

Bérard, S., Bergeron, A., Chauve, C., and Paul, C. (2007). Perfect sorting by reversals is not always difficult. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, 4(1):4–16.

Bérard, S., Chateau, A., Chauve, C., Paul, C., and Tannier, E. (2009). Computation of perfect DCJ rearrangement scenarios with linear and circular chromosomes. *J. Comput. Biol.*, 16(10):1287–1309.

Bergeron, A. (2005). A very elementary presentation of the Hannenhalli-Pevzner theory. *Discrete Appl. Math.*, 146(2):134–145.

Bergeron, A., Heber, S., and Stoye, J. (2002). Common intervals and sorting by reversals: a marriage of necessity. In *Proc. ECCB*, pages 54–63.

Bergeron, A., Mixtacki, J., and Stoye, J. (2006a). On sorting by translocations. *J. Comput. Biol.*, 13(2):567–578.

Bergeron, A., Mixtacki, J., and Stoye, J. (2006b). A unifying view of genome rearrangements. In *Proc. WABI*, pages 163–173.

Bergeron, A., Mixtacki, J., and Stoye, J. (2008). HP distance via double cut and join distance. In *Proc. CPM*, pages 56–68.

Bergeron, A., Mixtacki, J., and Stoye, J. (2009). A new linear time algorithm to compute the genomic distance via the double cut and join distance. *Theor. Comput. Sci.*, 410(51):5300–5316.

Bergeron, A. and Mixtacki, J. and Stoye, J. (2004). Reversal distance without hurdles and fortresses. In *Proc. CPM*, pages 388–399. Springer.

Berman, P. and Hannenhalli, S. (1996). Fast sorting by reversal. In *Proc. CPM*, pages 168–185.

Berman, P., Hannenhalli, S., and Karpinski, M. (2002). 1.375-approximation algorithm for sorting by reversals. In *Proc. ESA*, pages 401–408. Springer.

Berman, P. and Karpinski, M. (1999). On some tighter inapproximability results (extended abstract). In *Proc. ICALP*, pages 705–705. Springer.

Bernt, M., Merkle, D., and Middendorf, M. (2007). Using median sets for inferring phylogenetic trees. *Bioinformatics*, 23(2):e129.

Biedl, T. C. (2001). Linear reductions of maximum matching. In *Proc. SODA*, pages 825–826.

Bininda-Emonds, O. R. (2004). *Phylogenetic supertrees: combining information to reveal the tree of life*, volume 4. Springer.

Blanchette, M., Bourque, G., and Sankoff, D. (1997). Breakpoint phylogenies. In *Genome Inform. Ser. Workshop Genome Inform.*, pages 25–34.

Blin, G., Chauve, C., and Fertin, G. (2004). The breakpoint distance for signed sequences. In *Proc. CompBioNets*, pages 3–16.

Bourque, G. and Pevzner, P. (2002). Genome-scale evolution: reconstructing gene orders in the ancestral species. *Genome Res.*, 12(1):26.

Bourque, G., Zdobnov, E., Bork, P., Pevzner, P., and Tesler, G. (2005). Comparative architectures of mammalian and chicken genomes reveal highly variable rates of genomic rearrangements across different lineages. *Genome Res.*, 15(1):98.

Bruno, W. J., Socci, N. D., and Halpern, A. L. (2000). Weighted neighbor joining: a likelihood-based approach to distance-based phylogeny reconstruction. *Mol. Biol. Evol.*, 17(1):189–197.

Bryant, D. (1998). The complexity of the breakpoint median problem. Technical Report CRM-2579, Centre de Recherches Mathematiques, Universite de Montreal.

Bryant, D. (2000). The complexity of calculating exemplar distances. In Sankoff and Nadeau (2000), pages 207–212.

Bryant, D. (2004). A lower bound for the breakpoint phylogeny problem. *J. Discrete Algorithms*, 2(2):229–255.

Bulteau, L., Fertin, G., and Rusu, I. (2012a). Pancake flipping is hard. In *Proc. MFCS*, pages 247–258. Springer.

Bulteau, L., Fertin, G., and Rusu, I. (2012b). Sorting by transpositions is difficult. *SIAM J. Discrete Math.*, 26(3):1148–1180.

Caprara, A. (1997). Sorting by reversals is difficult. In *Proc. RECOMB*, page 83. ACM.

Caprara, A. (1999). Sorting permutations by reversals and Eulerian cycle decompositions. *SIAM J. Discrete Math.*, 12(1):91–110.

Caprara, A. (2001). On the practical solution of the reversal median problem. In *Proc. WABI*, pages 238–251.

Caprara, A. (2002). Additive bounding, worst-case analysis, and the breakpoint median problem. *SIAM J. Optimiz.*, 13(2):508–519.

Caprara, A. (2003). The reversal median problem. *INFORMS J. Comput.*, 15(1):93.

Chauve, C. and Tannier, E. (2008). A methodological framework for the reconstruction of contiguous regions of ancestral genomes and its application to mammalian genomes. *PLoS Comput. Biol.*, 4(11):e1000234.

Chen, Z., Fu, B., and Zhu, B. (2006). The approximability of the exemplar breakpoint distance problem. In *Proc. AAIM*, pages 291–302.

Chitturi, B., Fahle, W., Meng, Z., Morales, L., Shields, C., Sudborough, I., and Voit, W. (2009). An $(18/11)n$ upper bound for sorting by prefix reversals. *Theor. Comput. Sci.*, 410(36):3372–3390.

Chor, B. and Tuller, T. (2005). Maximum likelihood of evolutionary trees: hardness and approximation. *Bioinformatics*, 21(suppl 1):i97–i106.

Chor, B. and Tuller, T. (2006). Finding a maximum likelihood tree is hard. *J. ACM*, 53(5):722–744.

Christie, D. A. (1996). Sorting permutations by block-interchanges. *Inf. Process. Lett.*, 60(4):165–169.

Chrobak, M., Kolman, P., and Sgall, J. (2005). The greedy algorithm for the minimum common string partition problem. *ACM T. Algorithms*, 1(2):350–366.

Chrobak, M., Szymacha, T., and Krawczyk, A. (1990). A data structure useful for finding hamiltonian cycles. *Theor. Comput. Sci.*, 71(3):419–424.

Cohen, D. and Blum, M. (1995). On the problem of sorting burnt pancakes. *Discrete Appl. Math.*, 61(2):105–120.

Cosner, M. (1993). *Phylogenetic and molecular evolutionary studies of chloroplast DNA variation in the Campanulaceae*. PhD thesis, Ohio State University.

Cosner, M., Jansen, R., Moret, B., Raubeson, L., Wang, L., Warnow, T., and Wyman, S. (2000). An empirical comparison of phylogenetic methods on chloroplast gene order data in Campanulaceae. In Sankoff and Nadeau (2000), pages 99–121.

Cui, Y., Wang, L., Zhu, D., and Liu, X. (2008). A $(1.5 + \varepsilon)$-approximation algorithm for unsigned translocation distance. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, 5(1):56–66.

Day, W. (1983). Computationally difficult parsimony problems in phylogenetic systematics. *J. Theor. Biol.*, 103(3):429–438.

Dees, J. (2009). *Simultaneous Matchings in Dynamic Graphs*. Student research project, Universität Karlsruhe.

Desper, R. and Gascuel, O. (2004). Theoretical foundation of the balanced minimum evolution method of phylogenetic inference and its relationship to weighted least-squares tree fitting. *Mol. Biol. Evol.*, 21(3):587–598.

Doyon, J.-P., Ranwez, V., Daubin, V., and Berry, V. (2011). Models, algorithms and programs for phylogeny reconciliation. *Brief. Bioinform.*, 12(5):392–400.

El-Mabrouk, N., Bryant, D., and Sankoff, D. (1999). Reconstructing the pre-doubling genome. In *Proc. RECOMB*, pages 154–163. ACM.

El-Mabrouk, N. and Nadeau, J. and Sankoff, D. (1998). Genome halving. In *Proc. CPM*, pages 235–250. Springer.

El-Mabrouk, N. and Sankoff, D. (1999a). Hybridization and genome rearrangement. In *Proc. CPM*, pages 78–87. Springer.

El-Mabrouk, N. and Sankoff, D. (1999b). On the reconstruction of ancient doubled circular genomes using minimum reversals. *Genome Inform. Ser.*, pages 83–93.

El-Mabrouk, N. and Sankoff, D. (2003). The reconstruction of doubled genomes. *SIAM J. Comput.*, 32(3):754–792.

Elias, I. and Hartman, T. (2006). A 1.375-approximation algorithm for sorting by transpositions. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, 3(4):369–379.

Elias, I. and Tuller, T. (2007). Reconstruction of ancestral genomic sequences using likelihood. *J. Comput. Biol.*, 14(2):216–237.

Erdös, P. L., Soukup, L., and Stoye, J. (2011). Balanced vertices in trees and a simpler algorithm to compute the genomic distance. *Appl. Math. Lett.*, 24(1):82–86.

Eriksen, N. (2002). $(1 + \varepsilon)$-approximation of sorting by reversals and transpositions. *Theor. Comput. Sci.*, 289(1):517–529.

Eriksen, N. (2007). Reversal and transposition medians. *Theor. Comput. Sci.*, 374(1-3):111–126.

Eriksen, N. (2009). Median clouds and a fast transposition median solver. In *Proc. FPSAC*, pages 373–384. DMTCS Proceedings.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.*, 17(6):368–376.

Felsenstein, J. (2002). *PHYLIP (Phylogeny Inference Package) version 3.6 a3.*

Feng, J. and Zhu, D. (2007). Faster algorithms for sorting by transpositions and sorting by block interchanges. *ACM T. Algorithms*, 3(3).

Fertin, G., Labarre, A., and Rusu, I. (2009). *Combinatorics of genome rearrangements.* The MIT Press.

Figeac, M. and Varré, J.-S. (2004). Sorting by reversals with common intervals. In *Proc. WABI*, pages 26–37.

Fitch, W. and Margoliash, E. (1967). Construction of phylogenetic trees. *Science*, 155(760):279–284.

Foulds, L. and Graham, R. (1982). The Steiner problem in phylogeny is NP-complete. *Adv. Appl. Math.*, 3(1):43–49.

Gabow, H. (1973). *Implementation of algorithms for maximum matching on nonbipartite graphs.* PhD thesis, Stanford University.

Gabow, H. (1990). Data structures for weighted matching and nearest common ancestors with linking. In *Proc. SODA*, pages 434–443. Society for Industrial and Applied Mathematics.

Gabow, H. and Tarjan, R. (1991). Faster scaling algorithms for general graph matching problems. *J. ACM*, 38(4):815–853.

Garey, M. R. and Johnson, D. S. (1979). *Computers and Intractability: A Guide to the Theory of NP-Completeness.* W. H. Freeman.

Gavranović, H., Chauve, C., Salse, J., and Tannier, E. (2011). Mapping ancestral genomes with massive gene loss: A matrix sandwich problem. *Bioinformatics*, 27(13):i257–i265.

Gavranovic, H. and Tannier, E. (2010). Guided genome halving: provably optimal solutions provide good insights into the preduplication ancestral genome of *Saccharomyces cerevisiae*. In *Proc. Pac. Symp. Biocomp.*, volume 15, pages 21–30.

Goldstein, A., Kolman, P., and Zheng, J. (2005). Minimum common string partition problem: Hardness and approximations. *Electr. J. Comb.*, 12.

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of phyml 3.0. *Syst. Biol.*, 59(3):307–321.

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

Han, Y. (2006). Improving the efficiency of sorting by reversals. In Arabnia, H. R. and Valafar, H., editors, *Proc. Pac. Symp. Biocomp.*, pages 406–409. CSREA Press.

Hannenhalli, S. and Pevzner, P. (1999). Transforming cabbage into turnip: polynomial algorithm for sorting signed permutations by reversals. *J. ACM*, 46(1):1–27.

Hannenhalli, S. and Pevzner, P. A. (1995). Transforming men into mice (polynomial algorithm for genomic distance problem). In *Proc. FOCS*, pages 581–592.

Hasegawa, M., Kishino, H., and Yano, T. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial dna. *J. Mol. Evol.*, 22(2):160–174.

Heydari, M. H. and Sudborough, I. H. (1997). On the diameter of the pancake network. *J. Algorithm.*, 25(1):67–94.

Jackson, B., Schnable, P., and Aluru, S. (2007). Consensus genetic maps as median orders from inconsistent sources. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, pages 161–171.

Jean, G. and Nikolski, M. (2007). Genome rearrangements: a correct algorithm for optimal capping. *Inf. Process. Lett.*, 104(1):14–20.

Jukes, T. and Cantor, C. (1969). Evolution of protein molecules. *Mammalian protein metabolism*, 3:121–132.

Kaplan, H., Shamir, R., and Tarjan, R. E. (1999). A faster and simpler algorithm for sorting signed permutations by reversals. *SIAM J. Comput.*, 29(3):880–892.

Kaplan, H. and Verbin, E. (2005). Sorting signed permutations by reversals, revisited. *J. Comput. Syst. Sci.*, 70(3):321–341.

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.*, 16(2):111–120.

Knuth, D. (1973). *The art of computer programming, Vol. 3.* Addison-Wesley, Reading, MA.

Kolman, P. and Walen, T. (2007). Reversal distance for strings with duplicates: Linear time approximation using hitting set. *Electr. J. Comb.*, 14(1).

Kováč, J., Braga, M. D. V., and Stoye, J. (2010). The problem of chromosome reincorporation in DCJ sorting and halving. In *Proc. RECOMB-CG*, pages 13–24.

Kováč, J., Brejová, B., and Vinař, T. (2011a). A practical algorithm for ancestral rearrangement reconstruction. In Przytycka and Sagot (2011), pages 163–174.

Kováč, J., Warren, R., Braga, M. D. V., and Stoye, J. (2011b). Restricted DCJ model: Rearrangement problems with chromosome reincorporation. *J. Comput. Biol.*, 18(9):1231–1241.

Lander, E., Linton, L., Birren, B., Nusbaum, C., Zody, M., Baldwin, J. and Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.

Larget, B., Kadane, J., and Simon, D. (2005). A Bayesian approach to the estimation of ancestral genome arrangements. *Mol. Phylogenet. Evol.*, 36(2):214–223.

Lawler, E. (1976). *Combinatorial optimization: networks and matroids.* Holt, Rinehart and Winston.

Lin, Y. and Moret, B. M. (2008). Estimating true evolutionary distances under the dcj model. *Bioinformatics*, 24(13):i114–i122.

Lin, Y., Rajan, V., Swenson, K. M., and Moret, B. M. (2010). Estimating true evolutionary distances under rearrangements, duplications, and losses. *BMC Bioinformatics*, 11(Suppl 1):S54.

Ma, J. and Zhang, L., Suh, B., Raney, B.J. and Burhans, R., Kent, W.J. and Blanchette, M., Haussler, D., and Miller, W. (2006). Reconstructing contiguous regions of an ancestral genome. *Genome Res.*, 16(12):1557.

Maddison, D. and Maddison, W. (2000). *MacClade 4.0: Analysis of phylogeny and character evolution.* Sinauer Associates, Sunderland, Mass.

Maddison, W. and Maddison, D. (2004). Mesquite: a modular system for evolutionary analysis.

Maddison, W. P. (1997). Gene trees in species trees. *Syst. Biol.*, 46(3):523–536.

Micali, S. and Vazirani, V. V. (1980). An $O(\sqrt{|V|}|E|)$ algorithm for finding maximum matching in general graphs. In *Proc. FOCS*, pages 17–27. IEEE Computer Society.

Mixtacki, J. (2008). Genome halving under DCJ revisited. *Computing and Combinatorics*, pages 276–286.

Moret, B., Tang, J. and Wang, L., and Warnow, T. (2002a). Steps toward accurate reconstructions of phylogenies from gene-order data. *J. Comput. Syst. Sci.*, 65(3):508–525.

Moret, B. and Tang, J. and Warnow, T. (2005). Reconstructing phylogenies from gene-content and gene-order data. *Mathematics of Evolution and Phylogeny*, pages 321–352.

Moret, B., Wang, L., Warnow, T., and Wyman, S. (2001a). New approaches for reconstructing phylogenies from gene order data. *Bioinformatics*, 17(S1):S165.

Moret, B. M. (2005). Computational challenges from the tree of life. In *Proc. ALENEX*, pages 3–16. SIAM.

Moret, B. M. E., Siepel, A. C., Tang, J., and Liu, T. (2002b). Inversion medians outperform breakpoint medians in phylogeny reconstruction from gene-order data. In *Proc. WABI*, pages 521–536.

Moret, B. M. E., Wyman, S. K., Bader, D. A., Warnow, T., and Yan, M. (2001b). A new implementation and detailed study of breakpoint analysis. In *Proc. Pac. Symp. Biocomp.*, pages 583–594.

Nei, M. and Kumar, S. (2000). *Molecular evolution and phylogenetics*. Oxford University Press, USA.

Ozery-Flato, M. and Shamir, R. (2003). Two notes on genome rearrangement. *J. Bioinform. Comput. Biol.*, 1(1):71–94.

Ozery-Flato, M. and Shamir, R. (2006). An $O(n^{3/2}\sqrt{\log n})$ algorithm for sorting by reciprocal translocations. In Lewenstein, M. and Valiente, G., editors, *Proc. CPM*, volume 4009 of *LNCS*, pages 258–269. Springer.

Ozery-Flato, M. and Shamir, R. (2007). Rearrangements in genomes with centromeres part i: Translocations. In *Proc. RECOMB*, pages 339–353.

Page, R. D. and Charleston, M. A. (1997). From gene to organismal phylogeny: reconciled trees and the gene tree/species tree problem. *Mol. Phylogenet. Evol.*, 7(2):231–240.

Pe'er, I. and Shamir, R. (1998). The median problems for breakpoints are NP-complete. *Electronic Colloquium on Computational Complexity (ECCC)*, 5(71).

Pe'er, I. and Shamir, R. (2000). Approximation algorithms for the median problem in the breakpoint model. In Sankoff and Nadeau (2000), pages 225–241.

Plesník, J. (1979). The NP-completeness of the hamiltonian cycle problem in planar digraphs with degree bound two. *Inf. Process. Lett.*, 8(4):199–201.

Przytycka, T. M. and Sagot, M.-F., editors (2011). *Proc. WABI*, volume 6833 of *Lect. Notes in Computer Sci.* Springer.

Radcliffe, A., Scott, A., and Wilmer, E. (2006). Reversals and transpositions over finite alphabets. *SIAM J. Discrete Math.*, 19(1):224.

Rajan, V., Xu, A. W., Lin, Y., Swenson, K. M., and Moret, B. M. E. (2010). Heuristics for the inversion median problem. *BMC Bioinformatics*, 11(S-1):30.

Rokas, A. and Holland, P. W. (2000). Rare genomic changes as a tool for phylogenetics. *Trends Ecol. Evol.*, 15(11):454–459.

Ronquist, F. and Huelsenbeck, J. (2003a). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, 19(12):1572.

Ronquist, F. and Huelsenbeck, J. (2003b). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, 19(12):1572.

Sankoff, D. (1992). Edit distances for genome comparisons based on non-local operations. In *Proc. CPM*, pages 121–135.

Sankoff, D. and Blanchette, M. (1997). The median problem for breakpoints in comparative genomics. In *Proc. COCOON*, pages 251–264.

Sankoff, D. and Blanchette, M. (1998). Multiple genome rearrangement and breakpoint phylogeny. *J. Comput. Biol.*, 5(3):555–570.

Sankoff, D., Bryant, D., Deneault, M., Lang, B., and Burger, G. (2000). Early eukaryote evolution based on mitochondrial gene order breakpoints. *J. Comput. Biol.*, 7(3-4):521–535.

Sankoff, D., Cedergren, R. J., and Lapalme, G. (1976). Frequency of insertion-deletion, transversion, and transition in the evolution of 5s ribosomal rna. *J. Mol. Evol.*, 7(2):133–149.

Sankoff, D. and Nadeau, J. H., editors (2000). *Comparative Genomics: Empirical and Analytical Approaches to Gene Order Dynamics, Map Alignment, and the Evolution of Gene Families*, volume 1 of *Computational Biology Series*. Kluwer Academic Publishers.

Sankoff, D., Sundaram, G., and Kececioglu, J. D. (1996). Steiner points in the space of genome rearrangements. *Int. J. Found. Comput. Sci.*, 7(1):1–9.

Seoighe, C. and Wolfe, K. H. (1998). Extent of genomic rearrangement after genome duplication in yeast. *Proc. Natl. Acad. Sci.*, 95(8):4447–4452.

Siepel, A. C. and Moret, B. M. E. (2001). Finding an optimal inversion median: Experimental results. In *Proc. WABI*, pages 189–203.

Stamatakis, A. (2006). Raxml-vi-hpc: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22(21):2688–2690.

Stamatakis, A., Ludwig, T., and Meier, H. (2005). RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics*, 21(4):456–463.

Studier, J. A., Keppler, K. J., et al. (1988). A note on the neighbor-joining algorithm of saitou and nei. *Mol. Biol. Evol.*, 5(6):729–731.

Swenson, K. M., Badr, G., and Sankoff, D. (2011). Listing all sorting reversals in quadratic time. *Algorithm. Mol. Biol.*, 6:11.

Swenson, K. M. and Moret, B. M. E. (2009). Inversion-based genomic signatures. *BMC Bioinformatics*, 10(S-1).

Swenson, K. M., Rajan, V., Lin, Y., and Moret, B. M. E. (2010). Sorting signed permutations by inversions in $O(n \log n)$ time. *J. Comput. Biol.*, 17(3):489–501.

Swofford, D. (2003). *PAUP*. Phylogenetic analysis using parsimony (*and other methods). Version 4.*

Tamura, K. (1992). Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+ C-content biases. *Mol. Biol. Evol.*, 9(4):678.

Tamura, K. and Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.*, 10(3):512.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., and Kumar, S. (2011). MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.*, 28(10):2731–2739.

Tang, J. and Moret, B. (2005). Linear programming for phylogenetic reconstruction based on gene rearrangements. In *Proc. CPM*, pages 406–416. Springer.

Tang, J. and Wang, L.-S. (2005). Improving genome rearrangement phylogeny using sequence-style parsimony. In *Proc. BIBE*, pages 137–144. IEEE.

Tannier, E., Bergeron, A., and Sagot, M.-F. (2007). Advances on sorting by reversals. *Discrete Appl. Math.*, 155(6-7):881–888.

Tannier, E., Zheng, C., and Sankoff, D. (2009). Multichromosomal median and halving problems under different genomic distances. *BMC Bioinformatics*, 10.

Tesler, G. (2002). Efficient algorithms for multichromosomal genome rearrangements. *J. Comput. Syst. Sci.*, 65(3):587–609.

Valach, M., Farkaš, Z., Fričová, D., Kováč, J., Brejová, B., Vinař, T., Pfeiffer, I., Kucsera, J., Tomáška, Ľ., Lang, B. F., and Nosek, J. (2011). Evolution of linear chromosomes and multipartite genomes in yeast mitochondria. *Nucleic Acids Res.*, 39(10):4202–4219.

Wang, L.-S., Jansen, R. K., Moret, B. M. E., Raubeson, L. A., and Warnow, T. (2002). Fast phylogenetic methods for the analysis of genome rearrangement data: An empirical study. In *Proc. Pac. Symp. Biocomp.*, pages 524–535.

Wang, L.-S. and Warnow, T. (2001). Estimating true evolutionary distances between genomes. In *Proc. STOC*, pages 637–646. ACM.

Warnow, T. (2006). Disk covering methods: Improving the accuracy and speed of large-scale phylogenetic analyses. In *Handbook of computational molecular biology*, volume 9. CRC Press.

Warren, R. and Sankoff, D. (2009a). Genome aliquoting with double cut and join. *BMC Bioinformatics*, 10(Suppl 1):S2.

Warren, R. and Sankoff, D. (2009b). Genome halving with double cut and join. *J. Bioinform. Comput. Biol.*, 7(2):357–371.

Warren, R. and Sankoff, D. (2011). Genome aliquoting revisited. *J. Comput. Biol.*, 18(9):1065–1075.

Wheeler, T. J. (2009). Large-scale neighbor-joining with NINJA. In Salzberg, S. and Warnow, T., editors, *Proc. WABI*, volume 5724 of *Lect. Notes in Computer Sci.*, pages 375–389. Springer.

Xu, A. W. (2009). DCJ median problems on linear multichromosomal genomes: Graph representation and fast exact solutions. In Ciccarelli, F. and Miklós, I., editors, *Proc. RECOMB-CG*, volume 5817 of *Lect. Notes in Computer Sci.*, pages 70–83. Springer.

Xu, A. W. and Moret, B. M. E. (2011). GASTS: parsimony scoring under rearrangements. In Przytycka and Sagot (2011), pages 351–363.

Xu, A. W. and Sankoff, D. (2008). Decompositions of multiple breakpoint graphs and rapid exact solutions to the median problem. In *Proc. WABI*, pages 25–37.

Yancopoulos, S., Attie, O., and Friedberg, R. (2005). Efficient sorting of genomic permutations by translocation, inversion and block interchange. *Bioinformatics*, 21(16):3340–3346.

Zhang, M., Arndt, W., and Tang, J. (2008). A branch-and-bound method for the multichromosomal reversal median problem. In Crandall, K. A. and Lagergren, J., editors, *Proc. WABI*, volume 5251 of *Lect. Notes in Computer Sci.*, pages 14–24. Springer.

Zhang, M., Arndt, W., and Tang, J. (2009). An exact solver for the DCJ median problem. In Altman, R. B., Dunker, A. K., Hunter, L., Murray, T., and Klein, T. E., editors, *Proc. Pac. Symp. Biocomp.*, pages 138–149.

Zheng, C., Zhu, Q., Adam, Z., and Sankoff, D. (2008). Guided genome halving: hardness, heuristics and the history of the Hemiascomycetes. In *Proc. ISMB*, pages 96–104.

Zheng, C., Zhu, Q., and Sankoff, D. (2006). Genome halving with an outgroup. *Evol. Bioinform. Online*, 2:295.

Zhu, D. and Wang, L. (2006). On the complexity of unsigned translocation distance. *Theor. Comput. Sci.*, 352(1-3):322–328.

Abstract

In this thesis, we study several algorithmic problems from the field of genome rearrangements. During evolution, genomes undergo large-scale mutations. A segment of DNA can get reversed, moved to another position, or even another chromosome.

If we compare genomes of related extant species, we can find long conserved regions of DNA (such as genes) which are very similar sequentially, however their order is different. This motivates the following biological problems which also pose intriguing challenges for computer science:

- How related are the two given organisms?
- How did their ancestor look like?
- More generally: If we know gene orders of multiple species and their phylogenetic tree, how did the ancestral genomes look like?
- If we know just the gene orders of multiple species, what is their phylogenetic tree?

We formulate these questions as optimization problems: assuming a genome model with a fixed set of allowed rearrangement operations, we can define distance between two genomes as the minimum number of rearrangements necessary to transform one genome into the other. The problems of reconstructing the evolutionary history and the phylogenetic tree of given species is also formulated using the parsimony criterion: we search a phylogenetic tree and ancestral genomes which minimize the total number of rearrangement mutations in the evolutionary history.

In this thesis, we are interested in both theoretical and practical problems in genome rearrangements. We propose a new approach to ancestral genome reconstruction and we implement one of the first practical tools applicable to analysis of real datasets spanning a complex phylogeny and accommodating a variety of genome architectures. We demonstrate the accuracy of our program on the well-studied dataset of *Campanulaceae* chloroplast genomes, and apply it to the reconstruction of rearrangement histories of newly sequenced mitochondrial genomes of pathogenic yeasts from *Hemiascomycetes* clade.

We revisit the restricted DCJ model by Yancopoulos et al. We propose an $O(n \log n)$ time algorithm for sorting in this model, thus improving on the existing quadratic algorithm, and develop a new linear time algorithm for genome halving.

Our main results concern several open problems in the breakpoint model. We give an $O(n\sqrt{n})$ algorithm for the median problem improving on the existing cubic algorithm. Furthermore, we show that the problem is equivalent to finding maximum matching. Thus, any improvement to our solution would imply a better algorithm for the maximum matching, which has been an open problem for more than 30 years. We also prove that the more general small phylogeny problem is NP-hard. Surprisingly, we show that it is NP-hard (even APX-hard) already for four species. In other words, while finding an ancestor of three species is easy, finding two ancestors of four species is already hard. We thereby solve two open problems from the monograph by Fertin et al.: *Combinatorics of genome rearrangements.*

Vlastné publikácie autora

- Kováč, J., Braga, M. D. V., and Stoye, J. (2010). The problem of chromosome reincorporation in DCJ sorting and halving. In Tannier, E., editor, Science, RECOMB-CG, volume 6398 of Lecture Notes in Computer pages 13–24. Springer.

  Citácie:

  - Lv, J., Havlak, P., and Putnam, N. H. (2011). Constraints on genes shape long-term conservation of macro-synteny in metazoan genomes. BMC bioinformatics, 12(Suppl 9), S11.

- Kováč, J., Brejová, B., and Vinař, T. (2011a). A practical algorithm for ancestral rearrangement reconstruction. In Przytycka, T. M. and Sagot, M.-F., editors, *Algorithms in Bioinformatics - 11th International Workshop, WABI 2011, Saarbrücken, Germany, September 5-7, 2011. Proceedings*, volume 6833 of *Lecture Notes in Computer Science*, pages 163–174. Springer.

  Citácie:

  - Holloway, P., Swenson, K., Ardell, D., and El-Mabrouk, N. (2012, January). Evolution of genome organization by duplication and loss: an alignment approach. In Research in Computational Molecular Biology (pp. 94-112). Springer Berlin Heidelberg.
  - Zheng, C., Albert, V. A., Lyons, E., and Sankoff, D. (2012). Ancient angiosperm hexaploidy meets gene order reconstruction of the eudicot ancestor. In Second IEEE International Conference on Computational Advances in Bio and Medical Sciences (ICCABS).
  - Holloway, P., Swenson, K., Ardell, D., and El-Mabrouk, N. (2012, January). Evolution of genome organization by duplication and loss: an alignment approach. In Research in Computational Molecular Biology (pp. 94-112). Springer Berlin Heidelberg.
  - Holloway, P., Swenson, K., Ardell, D., and El-Mabrouk, N. (2013). Ancestral genome organization: an alignment approach. *J. Comput. Biol.* 20(4):280-95.

- Kováč, J., Warren, R., Braga, M. D. V., and Stoye, J. (2011b). Restricted DCJ model: Rearrangement problems with chromosome reincorporation. *J. Comput. Biol.,* 18(9):1231–1241.

  Citácie:

  - Hilker, R., Sickinger, C., Pedersen, C. N., and Stoye, J. (2012). UniMoG—a unifying framework for genomic distance calculation and sorting based on DCJ. *Bioinformatics,* 28(19), 2509-2511.
  - da Silva, P. H., Machado, R., Dantas, S., and Braga, M. D. (2012). Restricted DCJ-indel model: sorting linear genomes with DCJ and indels. *BMC bioinformatics,* 13(Suppl 19), S14.
  - Thomas, A., Ouangraoua, A., and Varré, J. S. (2012). Tandem halving problems by DCJ. In *Algorithms in Bioinformatics* (pp. 417-429). Springer Berlin Heidelberg.

- Valach, M., Farkaš, Z., Fričová, D., Kováč, J., Brejová, B., Vinař, T., Pfeiffer, I., Kucsera, J., Tomáška, Ľ., Lang, B. F., and Nosek, J. (2011). Evolution of linear chromosomes and multipartite genomes in yeast mitochondria. *Nucleic acids research,* 39(10):4202–4219.

  Citácie:

- Kayal, E., Bentlage, B., Collins, A. G., Kayal, M., Pirro, S., and Lavrov, D. V. (2012). Evolution of linear mitochondrial genomes in medusozoan cnidarians. Genome biology and evolution, 4(1), 1-12.
- Smith, D. R., Kayal, E., Yanagihara, A. A., Collins, A. G., Pirro, S., and Keeling, P. J. (2012). First complete mitochondrial genome sequence from a box jellyfish reveals a highly fragmented linear architecture and insights into telomere evolution. Genome biology and evolution, 4(1), 52-58.
- Haas, B. J., Zeng, Q., Pearson, M. D., Cuomo, C. A., and Wortman, J. R. (2011). Approaches to fungal genome annotation. Mycology, 2(3), 118-141.
- Gaillardin, C., Neuvéglise, C., Kerscher, S., and Nicaud, J. M. (2012). Mitochondrial genomes of yeasts of the Yarrowia clade. FEMS yeast research, 12(3), 317-331.
- Eldarov, M. A., Mardanov, A. V., Beletsky, A. V., Ravin, N. V., and Skryabin, K. G. (2011). Complete sequence and analysis of the mitochondrial genome of the methylotrophic yeast Hansenula polymorpha DL-1. FEMS yeast research, 11(6), 464-472.
- Budd, A. (2012). Diversity of Genome Organisation. In Evolutionary Genomics (pp. 51-76). Humana Press.
- Burger, G., Jackson, C. J., and Waller, R. F. (2012). Unusual mitochondrial genomes and genes. In Organelle genetics (pp. 41-77). Springer Berlin Heidelberg.
- Jung, P. P., Friedrich, A., Reisser, C., Hou, J., and Schacherer, J. (2012). Mitochondrial Genome Evolution in a Single Protoploid Yeast Species. G3: Genes— Genomes— Genetics, 2(9), 1103-1111.
- Gioti, A., Nystedt, B., Li, W., Xu, J., Andersson, A., Averette, A. F., ... and Scheynius, A. (2013). Genomic Insights into the Atopic Eczema-Associated Skin Commensal Yeast Malassezia sympodialis. mBio, 4(1).
- Smith, D. R., and Keeling, P. J. (2013). Gene conversion shapes linear mitochondrial genome architecture. Genome biology and evolution.