6 Summary and conclusions

Family specific rates of protein evolution

Amino acid changing mutations in proteins are contstrained by purifying selection and accumulate at different rates. Rate variations can be disentangled between different effects: the variations among gene families, among lineages and among specific genes in specific lineages. This thesis rests on the key observation that one can assign a family-specific rate of evolution to individual protein families.

We analyze protein families comprising orthologous sequences from man, fugu, fly and worm. Scatter plots comparing evolutionary distances between proteins in one and the same family exhibit strong linear correlations. This suggests that a fast or slow mutation rate is very much an attribute of the gene family that we can observe in either genomic comparison. We measure family specific evolutionary rates by applying standard maximum likelihood tree estimation procedures.

The assumption that the number of mutations per time unit is constant, the so called molecular clock hypothesis, allows representing the evolution of the family by a rooted ultrametric phylogenetic tree. In such a tree the edge lengths are proportional to the estimated number of mutation events and the ultrametricity implies that all leaves are equally distant to the root. The fact that a protein's evolutionary rate differs for different lineages, i.e. that the molecular clock does not hold in general, is accounted for by reconstructing an additive tree rather than an ultrametric one. We apply both tree models to estimate a family specific evolutionary rate. Since the two tree models are nested models, we perform a likelihood ratio test and delineate a set of families well fitting the molecular clock assumption. The ultrametric tree model incorporates the Family Specific Rate as a scaling paramter. Its ML estimation yields almost the same results as the measure of the overall length of the additive ML tree when no prior assumption of a constant rate of evolution among lineages is made. The Family Specific Rates and a pregiven set of divergence times are therefore used to relate measures of amino acid replacements to historical times on a genome scale.

We analyze several of publicly available data sets with respect to the overall distribution of Family Specific Rates. First of all, we establish the relation of our rate measure to estimated numbers of nonsynonymous substitutions measured between two nematodes. Compared to numbers of nonsynonymous substitutions, the Family Specific Rate measure has the major effect of averaging over lineage specific rate variations.

Further, we establish relationships of Family Specific Rates to the essentiality and the dispensability of proteins in interaction networks that were assessed by RNAi knockout experiments and high throughput 2-hybrid systems in *C. elegans*, respectively. Purifying selection indeed acts stronger on essential genes than on nonessential ones. The observed relationships link experimental results that were obtained for the nematode to other eukaryotic model organisms.

The younger the faster

Interestingly, when grouping proteins according to their subcellular locale, we observe that extra-cellular proteins are fast evolving. Extra-cellular proteins were invented during metazoan evolution through gene duplication and domain shuffling events. We analyze the set of extra-cellular proteins and a specific large multigene family containing receptor tyrosine kinases in greater detail. From the observation that extra-cellularity is coupled to elevated evolutionary rates, we are motivated to set up a hypothesis: The evolutionary rate of a protein tends to be larger the more recently the protein emerged in evolution.

We investigate the hypothesis "the younger the faster" and perform PSI-BLAST searches of eukaryotic orthologous profiles against prokaryotic genomes. The experiment approves the dependancy of our means to detect homology by sequence comparison on evolutionary rates. The faster a protein evolves, the less we are able to trace its evolutionary origin.

The evolution of novel protein functions commonly relies on reusing and recombining already existing domains. The age of an orthologous family is reflected by the taxonomic distribution of proteins sharing the same domain architecture. Rate distributions of taxon-specific sets are in accordance with "the younger the faster".

We aim at placing an argument for the pertinence of assigning an age to an orthologous family by the taxonomic distribution of domain architectures. Duplication events predating the nematode-arthropode split gave rise to the emergence of new orthologous families. We analyze multigene families and compare duplication time points between taxon-specific sets. Duplication times are relatively small for multigene families with a "young" least common taxon and relatively large for multigene families with an "old" least common taxon.

Some duplication time estimates predating the earth's putative origin demand discussion. First of all, the overestimation of duplication times might be due to overestimated divergence times that were used to calibrate Family Specific Rates. Indeed, time estimates for the nematode-arthropode divergence that are obtained from molecular data vary by a factor of 2 and range from 550 to 1170 Millions of years. The divergence times that we used for calibration are at the upper limit.

Second, the relative rate test is a necessary but not a sufficient requirement for rate constancy to hold. Duplication times are overestimated when evolutionary rates along different lineages simultaneously decrease in time. The assertion also fits the observation that *Metazoa*- and *Bilateria*-specific sets lack families that evolve at small rates. Actually, the hypothesis "the younger the faster" implies such an assertion. The following interpretation for the "average protein" is plausible though speculative: Once a novel protein emerges, the tolerance against accepting mutations is large. Later, it occupies a specialized function and the rate decreases.

Trees on multiple genes

Correlations in evolutionary distances provide a rate independent signal of the underlying organismal phylogeny. We use these correlations as well as estimated evolutionary rates to propose two estimators for an organismal phylogenetic tree. We apply the estimators to our data set and add a selection criterion among families to focus on those that display rate constancy.

Selecting the families for rate constancy under a given tree model (our reference tree) would suggest that the computation of a new tree will only reinforce the reference tree. Interestingly, this is not the case. While the reference tree is ultrametric, the tree we compute from the ensemble of orthologous families is not ultrametric. Its inner edge is considerably shorter than in the reference tree. Thus, the data do not support the evolutionary times of the reference tree. Of course, additive, non-ultrametric edge lengths cannot be interpreted as historical times such that we are forced to assume that there was a change in evolutionary rate in either of the lineages of the tree. We therefore agree with the study of Peterson *et al.* [2004] that points to the possibility that evolutionary rates in vertebrate lineages were decreased with regard to rates in invertebrate lineages.

Bibliography

Abram, C. L. and Courtneidge, S. A. (2000) Src Family Tyrosine Kinases and Growth Factor Signaling. *Experimental Cell Research*, **254**, 1–13.

Adachi, J. and Hasegawa, M. (1996) MOLPHY: Programs for Molecular Phylogenetics, ver. 2.3. *Tokyo: Institute of Statistical Mathematics.*

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403–410.

Altschul, S. F., Madden, T. L., Schäffer, 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 Research*, **25**, 3389–3402.

Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M. J., Natale, D. A., O'Donovan, C., Redaschi, N. and Yeh, L. S. (2004) Uniprot: the universal protein knowledgebase. *Nucleic Acids Research*, **32**, D115–D119.

Bairoch, A. (2000) The ENZYME database in 2000. *Nucleic Acids Research*, **28**, 304–305.

Barry, D. and Hartigan, J. (1987) Asynchronous distance between homologous DNA sequences. *Biometrics*, **43(2)**, 261–276.

Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L. L., Studholme, D. J., Yeats, C. and Eddy, S. R. (2004) The Pfam protein families database. *Nucleic Acids Research*, **32**, D138–D141.

Bloom, J. D. and Adami, C. (2003) Apparent dependence of protein evolutionary rate on number of interactions is linked to biases in protein-protein interactions data sets. *BMC Evolutionary Biology*, **3**, 21.

Bloom, J. D. and Adami, C. (2004) Evolutionary rate depends on number of protein-protein interactions independently of gene expression level: response. *BMC Evolutionary Biology*, **4**, 14.

Brown, J. R., Douady, C. J., Italia, M. J., Marshall, W. E. and Stanhope, M. J. (2001) Universal trees based on large combined protein sequence data sets. *Nature Genetics*, **28**, 281–285.

Brown, M., Hughey, R., Krogh, A., Mian, I. and Haussler, D. (1993) Using Dirichlet Mixture Priors to Derive Hidden Markov Models for Protein Families. In Hunter, L., Searls, D. and Shavlik, J. (eds.), *ISMB 93: Proceedings of the First International Conference on Intelligent Systems for Molecular Biology*, pp. 47–55, Bethesda, MD, USA.

Brown, M. T. and Cooper, J. A. (1996) Regulation, Substrates and Functions of Src. *Biochimica et Biophysica Acta*, **1287**, 121–149.

Castillo-Davis, C. I., Kondrashov, F. A., Hartl, D. L. and Kulathinal, R. J. (2004) The functional genomic distribution of protein divergence in two animal phyla: coevolution, genomic conflict, and constraint. *Genome Research*, **14**, 802–811.

Chothia, C. (1994) Protein families in the metazoan genome. *Development*, **S**, 27–33.

Chothia, C., Gough, J., Vogel, C. and Teichmann, S. A. (2003) Evolution of the protein repertoire. *Science*, **300**, 1701–1703.

Coghlan, A. and Wolfe, K. H. (2002) Fourfold faster rate of genome rearrangement in nematodes than in Drosophila. *Genome Research*, **12**, 857–867.

Conte, L. L., Ailey, B., Hubbard, T. J. P., Brenner, S. E., Murzin, A. G. and Chothia, C. (2000) SCOP: a Structural Classification of Proteins database. *Nucleic Acids Research*, **28**, 257–259.

Cresko, W. A., Yan, Y. L., Baltrus, D. A., Amores, A., Singer, A., Rodriguez-Mari, A. and Postlethwait, J. H. (2003) Genome duplication, subfunction partitioning, and lineage divergence: Sox9 in stickleback and zebrafish. *Developmental Dynamics*, **228**, 480–489.

Cutter, A. D., Payseur, B. A., Salcedo, T., Estes, A. M., Good, J. M., Wood, E., Hartl, T., Maughan, H., Strempel, J., Wang, B., Bryan, A. C. and Dellos, M. (2003) Molecular correlates of genes exhibiting RNAi phenotypes in Caenorhabditis elegans. *Genome Research*, **13**, 2651–2657.

Davis, J. and Petrov, D. (2004) Preferential duplication of conserved proteins in eukaryotic genomes. *PLoS Biol*, **2**.

Dayhoff, M., Eck, R. and Park, C. (1972) A model of evolutionary change in proteins. In Dayhoff, M. O. (ed.), *Atlas of Protein Sequence and Structure*, volume 5, pp. 88–99, National Biomedical Research Foundation, Washington DC.

Dayhoff, M., Schwartz, R. and Orcutt, B. (1978) A model of evolutionary change in proteins. In Dayhoff, M. O. (ed.), *Atlas of Protein Sequence and Structure*, volume 5, pp. 345–352, National Biomedical Research Foundation, Washington DC.

Doolittle, R. F. (1995) The multiplicity of domains in proteins. *Annual Review of Biochemistry*, **64**, 287–314.

Eddy, S. R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.

Efron, B. and Tibshirani, R. (1993) *An Introduction to the Bootstrap.* Chapman and Hall, Cambridge.

Ewens, W. J. and Grant, G. R. (2001) *Statistical Methods in Bioinformatics. An Introduction.* Springer Verlag, New York, NY.

Felsenstein, J. (1973) Maximum-likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. *Systematic Zoology*, **22**, 240–249.

Felsenstein, J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, **17**, 368–376.

Felsenstein, J. (1983) Statistical inference of phylogenies. *J. Royal Statist. Soc. A*, **146**, 246–272.

Felsenstein, J. (1988) Phylogenies from molecular sequences: Inference and reliability. *Annual Review of Genomics and Human Genetics*, **22**, 521–565.

Felsenstein, J. (1993) *PHYLIP manual, version 3.5c.* Department of Genetics, University of Washington, Seattle.

Feng, D., Johnson, M. and Doolittle, R. (1985) Aligning amino acid sequences: comparison of commonly used methods. *Journal of Molecular Evolution*, **21**, 112–125.

Fitch, W. M. (1970) Distinguishing homologous from analogous proteins. *Systematic Zoology*, **19**, 99–113.

Fraser, H. B. and Hirsh, A. E. (2004) Evolutionary rate depends on number of protein-protein interactions independently of gene expression level. *BMC Evolutionary Biology*, **4**, 13.

Fraser, H. B., Wall, D. P. and Hirsh, A. E. (2003) A simple dependence between protein evolution rate and the number of protein-protein interactions. *BMC Evol Biol*, **3**, 11.

Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R. and Bairoch, A. (2003) ExPASy: the proteomics server for in–depth protein knowledge and analysis. *Nucleic Acids Research*, **31**, 3784–3788.

Geer, L. Y., Domrachev, M., Lipman, D. J. and Bryant, S. H. (2002) CDART: protein homology by domain architecture. *Genome Research*, **12**, 1619–1623.

Goldman, N. and Yang, Z. (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology and Evolution*, **11**, 725–736.

Graur, D. and Martin, W. (2004) Reading the entrails of chickens: molecular timescales of evolution and the illusion of precision. *Trends in Genetics*, **20**, 80–86.

Gribskov, M., McLachlan, A. D. and Eisenberg, D. (1987) Profile analysis: Detection of distantly related proteins. *Proceedings of the National Academy of Sciences USA*, **84**, 4355–4358.

Grishin, N. V., Wolf, Y. I. and Koonin, E. V. (2000) From complete genomes to measures of substitution rate variability within and between proteins. *Genome Research*, **10**, 991–1000.

Han, J. D., Bertin, N., Hao, T., Goldberg, D. S., Berriz, G. F., Zhang, L. V., Dupuy, D., Walhout, A. J., Cusick, M. E., Roth, F. P. and Vidal, M. (2004) Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature*, **430**, 88–93.

Hanks, S. K. and Hunter, T. (1995) Protein kinases 6. The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification. *The FASEB journal : official publication of the Federation of American Societies for Experimental Biology*, **9**, 576–596.

Harris, M. A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C., Richter, J., Rubin, G. M., Blake, J. A., Bult, C., Dolan, M., Drabkin, H., Eppig, J. T., Hill, D. P., Ni, L., Ringwald, M., Balakrishnan, R., Cherry, J. M., Christie, K. R., Costanzo, M. C., Dwight, S. S., Engel, S., Fisk, D. G., Hirschman, J. E., Hong, E. L., Nash, R. S., Sethuraman, A., Theesfeld, C. L., Botstein, D., Dolinski, K., Feierbach, B., Berardini, T., Mundodi, S., Rhee, S. Y., Apweiler, R., Barrell, D., Camon, E., Dimmer, E., Lee, V., Chisholm, R., Gaudet, P., Kibbe, W., Kishore, R., Schwarz, E. M., Sternberg, P., Gwinn, M., Hannick, L., Wortman, J., Berriman, M., Wood, V., de la Cruz, N., Tonellato, P., Jaiswal, P., Seigfried, T. and White, R. (2004) The gene ontology (go) database and informatics resource. *Nucleic Acids Research*, **32**, D258–D261.

Hedges, S. B. (2002) The origin and evolution of model organisms. *Nature Reviews Genetics*, **3**, 838–849.

Hedges, S. B. and Kumar, S. (2003) Genomic clocks and evolutionary timescales. *Trends in Genetics*, **19**, 200–206.

Henikoff, S. and Henikoff, J. G. (1992) Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences USA*, **89**, 10915–10919.

Higgins, D. G., Thompson, J. D. and Gibson, T. J. (1996) Using CLUSTAL for multiple sequence alignments. *Methods in Enzymology*, **266**, 383–402.

Hirsh, A. E. and Fraser, H. B. (2001) Protein dispensability and rate of evolution. *Nature*, **411**, 1046–1049.

Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., Durbin, R., Eyras, E., Gilbert, J., Hammond, M., Huminiecki, L., Kasprzyk, A., Lehvaslaiho, H., Lijnzaad, P., Melsopp, C., Mongin, E., Pettett, R., Pocock, M., Potter, S., Rust, A., Schmidt, E., Searle, S., Slater, G., Smith, J., Spooner, W., Stabenau, A., Stalker, J., Stupka, E., Ureta-Vidal, A., Vastrik, I. and Clamp, M. (2002) The Ensembl genome database project. *Nucleic Acids Research*, **30**, 38–41.

Hughes, A. L. (2002) Adaptive evolution after gene duplication. *Trends in Genetics*, **18**, 433–434.

Hughes, A. L., Hughes, M. K., Howell, C. Y. and Nei, M. (1994) Natural selection at the class II major histocompatibility complex loci of mammals. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences.*, **345**, 359–367.

Hurst, L. D. and Smith, N. G. (1999) Do essential genes evolve slowly? *Current Biology*, **9**, 747–750.

Jeong, H., Mason, S. P., Barabasi, A. L. and Oltvai, Z. N. (2001) Lethality and centrality in protein networks. *Nature*, **411**, 41–42.

Jordan, I. K., Rogozin, I. B., Wolf, Y. I. and Koonin, E. V. (2002) Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. *Genome Research*, **12**, 962–968.

Jordan, I. K., Wolf, Y. I. and Koonin, E. V. (2003) No simple dependence between protein evolution rate and the number of protein-protein interactions: only the most prolific interactors tend to evolve slowly. *BMC Evolutionary Biology*, **3**, 1.

Jordan, I. K., Wolf, Y. I. and Koonin, E. V. (2004) Duplicated genes evolve slower than singletons despite the initial rate increase. *BMC Evolutionary Biology*, **4**, 22.

Jukes, T. H. and Cantor, C. R. (1969) Evolution of protein molecules. In Munro, H. N. (ed.), *Mammalian Protein Metabolism*, volume 3, pp. 21–123, Academic Press, New York.

Kamath, R. S., Fraser, A. G., Dong, Y., Poulin, G., Durbin, R., Gotta, M., Kanapin, A., Bot, N. L., Moreno, S., Sohrmann, M., Welchman, D. P., Zipperlen, P. and Ahringer, J. (2003) Systematic functional analysis of the Caenorhabditis elegans genome using RNAi. *Nature*, **421**, 231–237.

Kimura, M. (1968) Evolutionary rate at the molecular level. *Nature*, **217**, 624–626.

Kimura, M. (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, **16**, 111–120.

Kondrashov, F. A., Rogozin, I. B., Wolf, Y. I. and Koonin, E. V. (2002) Selection in the evolution of gene duplications. *Genome Biology*, **3**, RESEARCH0008.

Koonin, E. V., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Krylov, D. M., Makarova, K. S., Mazumder, R., Mekhedov, S. L., Nikolskaya, A. N., Rao, B. S., Rogozin, I. B., Smirnov, S., Sorokin, A. V., Sverdlov, A. V., Vasudevan, S., Wolf, Y. I., Yin, J. J. and Natale, D. A. (2004) A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biology*, **5**, R7.

Krause, A., Stoye, J. and Vingron, M. (2005) Large scale hierarchical clustering of protein sequences. *BMC Bioinformatics*, **6**, 15.

Krause, A. and Vingron, M. (1998) A set-theoretic approach to database searching and clustering. *Bioinformatics*, **14**, 430–438.

Krogh, A., Brown, M., Mian, I. S., Sjolander, K. and Haussler, D. (1994) Hidden Markov models in computational biology. Applications to protein modeling. *Journal of Molecular Biology*, **235**, 1501–1531.

Kunin, V. and Ouzounis, C. A. (2003) GeneTRACE-reconstruction of gene content of ancestral species. *Bioinformatics*, **19**, 1412–1416.

Kunin, V., Pereira-Leal, J. B. and Ouzounis, C. A. (2004) Functional evolution of the yeast protein interaction network. *Molecular Biology and Evolution*, **21**, 1171–1176.

Lagarias, J., J. A. Reeds, M. H. W. and Wright, P. E. (1998) Convergence Properties of the Nelder–Mead Simplex Method in Low Dimensions. *SIAM Journal of Optimization*, **9**, 112–147.

Lake, J. A. (1994) Reconstructing evolutionary trees from DNA and protein sequences: Paralinear distances. *Proceedings of the National Academy of Sciences USA*, **91**, 1455–1459.

Lanave, C., Preparata, G., Saccone, C. and Serio, G. (1984) A new method for calculating evolutionary substitution rates. *Journal of Molecular Evolution*, **20**, 86–93.

Li, S., Armstrong, C. M., Bertin, N., Ge, H., Milstein, S., Boxem, M., Vidalain, P. O., Han, J. D., Chesneau, A., Hao, T., Goldberg, D. S., Li, N., Martinez, M., Rual, J. F., Lamesch, P., Xu, L., Tewari, M., Wong, S. L., Zhang, L. V., Berriz, G. F., Jacotot, L., Vaglio, P., Reboul, J., Hirozane-Kishikawa, T., Li, Q., Gabel, H. W., Elewa, A., Baumgartner, B., Rose, D. J., Yu, H., Bosak, S., Sequerra, R., Fraser, A., Mango, S. E., Saxton, W. M., Strome, S., Heuvel, S. V. D., Piano, F., Vandenhaute, J., Sardet, C., Gerstein, M., Doucette-Stamm, L., Gunsalus, K. C., Harper, J. W., Cusick, M. E., Roth, F. P., Hill, D. E. and Vidal, M. (2004) A map of the interactome network of the metazoan C. elegans. *Science*, **303**, 540–543.

Li, W.-H. (1983) Evolution of duplicate genes and pseudogenes. In Nei, M. and Koehn, R. (eds.), *Evolution of Genes and Proteins*, pp. 14–37, Sinauer, Sunderland MA.

Lindgren, W. (1993) *Statistical Theory*. Chapman and Hall, New York.

Lipman, D. J., Altschul, S. F. and Kececioglu, J. D. (1989) A tool for multiple sequence alignment. *Proceedings of the National Academy of Sciences USA*, **86**, 4412–4415.

Lipman, D. J., Souvorov, A., Koonin, E. V., Panchenko, A. R. and Tatusova, T. A. (2002) The relationship of protein conservation and sequence length. *BMC Evolutionary Biology*, **2**, 20.

Lockhart, P. J., Steel, M. A., Hendy, M. D. and Penny, D. (1994) Recovering evolutionary trees under a more realistic model of sequence evolution. *Molecular Biology and Evolution*, **11**, 605–612.

Lynch, M. and Conery, J. S. (2000) The evolutionary fate and consequences of duplicate genes. *Science*, **290**, 1151–1155.

Matys, V., Fricke, E., Geffers, R., Gossling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A. E., Kel-Margoulis, O. V., Kloos, D. U., Land, S., Lewicki-Potapov, B., Michael, H., Munch, R., Reuter, I., Rotert, S., Saxel, H., Scheer, M., Thiele, S. and Wingender, E. (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Research*, **31**, 374–378.

Meinel, T., Krause, A., Luz, H., Vingron, M. and Staub, E. (2005) The SYSTERS Protein Family Database in 2005. *Nucleic Acids Research*, in press.

Meinel, T., Vingron, M. and Krause, A. (2003) The SYSTERS protein family database: Taxon related protein family size distributions and singleton frequencies. In *Proceedings on the German Conference on Bioinformatics*.

Mott, R., Schultz, J., Bork, P. and Ponting, C. P. (2002) Predicting protein cellular localization using a domain projection method. *Genome Research*, **12**, 1168–1174.

Müller, T. (2001) *Modellierung von Proteinevolution, Dissertation Thesis* *http://www.iwr.uni-heidelberg.de/sfb359/PP/Preprint2001-30.pdf.gz*. IWR Schriftenreihe, Heidelberg.

Müller, T., Rahmann, S., Dandekar, T. and Wolf, M. (2003) Robust estimation of the phylogeny Chlorophyceae (Chlorophyta) based on profile distances. In *Proceedings on the German Conference on Bioinformatics*.

Müller, T., Rahmann, S., Dandekar, T. and Wolf, M. (2004) Accurate and robust phylogeny estimation based on profile distances: a study of the Chlorophyceae (Chlorophyta). *BMC Evol Biol*, **4**, 20.

Müller, T., Spang, R. and Vingron, M. (2002) A Comparison of Dayhoff's Estimator, the Resolvent Approach and a Maximum Likelihood Method. *Molecular Biology and Evolution*, **10**, 8–13.

Müller, T. and Vingron, M. (2000) Modeling Amino Acid Replacement. *Journal of Computational Biology*, **6**, 761–776.

Nair, R. and Rost, B. (2002a) Inferring sub-cellular localization through automated lexical analysis. *Bioinformatics*, **18**, S78–86.

Nair, R. and Rost, B. (2002b) *List of SWISS-PROT keywords with strong correlation to sub-cellular locaisation,*
http://cubic.bioc.columbia.edu/db/LOCkey/easy.html.

Needleman, S. B. and Wunsch, C. D. (1970) A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins. *Journal of Molecular Biology*, **48**, 443–453.

Nei, M., Xu, P. and Glazko, G. (2001) Estimation of divergence times from multiprotein sequences for a few mammalian species and several distantly related organisms. *Proceedings of the National Academy of Sciences USA*, **98**, 2497–502.

Nielsen, H., Engelbrecht, J., Brunak, S. and von Heijne, G. (1997) Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Engineering*, **10**, 1–6.

Nielsen, H. and Krogh, A. (1998) Prediction of signal peptides and signal anchors by a hidden Markov model.

Ohno, S. (1973) Ancient linkage groups and frozen accidents. *Nature*, **244**, 259–262.

Pace, N. R. (1997) A molecular view of microbial diversity and the biosphere. *Science*, **276**, 734–740.

Pearson, W. R. and Lipman, D. J. (1988) Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences USA*, **85**, 2444–2448.

Peterson, K. J., Lyons, J. B., Nowak, K. S., Takacs, C. M., Wargo, M. J. and McPeek, M. A. (2004) Estimating metazoan divergence times with a molecular clock. *Proceedings of the National Academy of Sciences USA*, **101**, 6536–6541.

Press, W. H., Flannery, B. P., Teukolsky, S. and Vetterling, W. T. (1999) *Numerical Recipes in C: The Art of Scientific Computing.* 2nd edition, Cambridge University Press, Cambridge.

Raes, J. and de Peer, Y. V. (2003) Gene duplication, the evolution of novel gene functions, and detecting functional divergence of duplicates in silico. *Appl Bioinformatics*, **2**, 91–101.

Remm, M., Storm, C. E. and Sonnhammer, E. L. (2001) Automatic clustering of orthologs and in–paralogs from pairwise species comparisons. *JMB*, **314**, 1041–1052.

Robinson, D. R., Wu, Y. M. and Lin, S. F. (2000) The protein tyrosine kinase family of the human genome. *Oncogene*, **19**, 5548–5557.

Saitou, N. and Nei, M. (1987) The neighbor–joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, **4**, 406–425.

Sarich, V. M. and Wilson, A. C. (1973) Generation time and genomic evolution in primates. *Science*, **179**, 1144–1147.

Schmidt, H. A., Strimmer, K., Vingron, M. and von Haeseler, A. (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics*, **18**, 502–504.

Schultz, J., Milpetz, F., Bork, P. and Ponting, C. P. (1998) SMART, a simple modular architecture research tool: Identification of signaling domains. *Proceedings of the National Academy of Sciences USA*, **95**, 5857–5864.

Smith, T. F. and Waterman, M. S. (1981) The identification of common molecular subsequences. *Journal of Molecular Biology*, **147**, 195–197.

Sonnhammer, E. L., Eddy, S. R., Birney, E., Bateman, A. and Durbin, R. (1998a) Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Research*, **26**, 320–322.

Sonnhammer, E. L., Eddy, S. R. and Durbin, R. (1997) Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins*, **28**, 405–420.

Sonnhammer, E. L. and Koonin, E. V. (2002) Orthology, paralogy and proposed classification for paralog subtypes. *Trends in Genetics*, **18**, 619–620.

Sonnhammer, E. L. L., von Heijne, G. and Krogh, A. (1998b) A hidden markov model for predicting transmembrane helices in protein sequences. In *Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology*, pp. 175–182.

Stein, L., Bao, Z., Blasiar, D., Blumenthal, T., Brent, M., Chen, N., Chinwalla, A., Clarke, L., Clee, C., Coghlan, A., Coulson, A., D'Eustachio, P., Fitch, D., Fulton, L., Fulton, R., Griffiths-Jones, S., Harris, T., Hillier, L., Kamath, R., Kuwabara, P., Mardis, E., Marra, M., Miner, T., Minx, P., Mullikin, J., Plumb, R., Rogers, J., Schein, J., Sohrmann, M., Spieth, J., Stajich, J., Wei, C., Willey, D., Wilson, R., Durbin, R. and Waterston, R. (2003) The Genome Sequence of Caenorhabditis briggsae: A Platform for Comparative Genomics. *PLoS Biology*, **1**.

Stoye, J. (1998) Multiple Sequence Alignment with the Divide-and-Conquer Method. *Gene*, **211**, 45–46.

Stoye, J., Evers, D. and Meyer, F. (1998) Rose: generating sequence families. *Bioinformatics*, **14**, 157–163.

Stoye, J., Moulton, V. and Dress, A. W. (1997) DCA: an efficient implementation of the divide–and–conquer approach to simultaneous multiple sequence alignment. *CABIOS*, **13**, 625–626.

Strimmer, K. and von Haeseler, A. (1996) Quartet Puzzling: A Quartet Maximum–likelihood Method for Reconstructing Tree Topologies. *Molecular Biology and Evolution*, **13**, 964–969.

Tatusov, R. L., Koonin, E. V. and Lipman, D. J. (1997) A Genomic Perspective on Protein Families. *Science*, **278**, 631–637.

Teichmann, S. A. (2002) The constraints protein-protein interactions place on sequence divergence. *Journal of Molecular Biology*, **324**, 399–407.

Thompson, J. D., Higgins, D. G. and Gibson, T. J. (1994) Clustal w: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, **22**, 4673–4680.

Vingron, M. and von Haeseler, A. (1997) Towards integration of Multiple Alignment and Phylogenetic Tree Construction. *Journal of Computational Biology*, **4**, 23–34.

Wagner, A. (2002) Selection and gene duplication: a view from the genome. *Genome Biology*, **3**, reviews1012.

Wang, D. Y., Kumar, S. and Hedges, S. B. (1999) Divergence time estimates for the early history of animal phyla and the origin of plants, animals and fungi. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, **266**, 162–171.

Winter, E. E., Goodstadt, L. and Ponting, C. P. (2004) Elevated rates of protein secretion, evolution, and disease among tissue-specific genes. *Genome Research*, **14**, 54–61.

Wolf, Y. I., Rogozin, I. B., Grishin, N. V., Tatusov, R. L. and Koonin, E. V. (2001) Genome trees constructed using five different approaches suggest new major bacterial clades. *BMC Evolutionary Biology*, **1**, 8.

Wootton, J. C. and Federhen, S. (1993) Statistics of Local Complexity in Amino Acid Sequences and Sequence Databases. *Computers and Biochemistry*, **17**, 149–163.

Wuchty, S. (2002) Interaction and Domain Networks in Yeast. *Proteomics*, **2**, 1715–1723.

Yang, Z. (1996) Maximum-Likelihood Models for Combined Analyses of Multiple Sequence Data. *Journal of Molecular Evolution*, **42**, 587–596.

Yang, Z. (1997) A program package for phylogenetic analysis by maximum likelihood. *CABIOS*, **13**, 431–439.

Zharkikh, A. (1994) Estimation of evolutionary distances between nucleotide sequences. *Journal of Molecular Evolution*, **39**, 315–329.

Zuckerkandl, E. and Pauling, L. (1962) Molecular disease, evolution and genetic heterogeneity. In Marsha, M. and Pullman, B. (eds.), *Horizons in Biochemistry*, pp. 189–225, Academic Press.

Zuckerkandl, E. and Pauling, L. (1965) Molecules as documents of evolution. *Journal of Theoretical Biology*, **8**, 357–366.